Содержание

Так называемый здравый смысл состоит в принципиальном игнорировании, замалчивании или высмеивании всего, что не соответствует традиционной концепции мира, будто бы полностью объясненного в девятнадцатом веке. А тем временем на каждом шагу можно столкнуться с явлениями, структуру которых не понимаешь и не поймешь без применения статистики. Например, чем объясняются прославленная duplicitas casuum врачей, или поведение толпы, или циклические флуктуации смысла снов, или случаи, происходящие с вертящимися столиками?

Станислав Лем
«Расследование»

От автора

Учебное пособие основано на авторском курсе лекций для студентов астрономического отделения первого курса физического факультета МГУ им. М. В. Ломоносова. Пособие может быть использовано студентами астрономических специальностей физических факультетов университетов, а также научными сотрудниками институтов, занимающихся исследованиями в области астрономии и физики космоса.

Пособие состоит из пятнадцати глав, приложения, списков использованной и рекомендуемой литературы и предметного указателя основных терминов. В главе 1 рассматриваются понятия «погрешность» и «ошибка измерения», описываются математические операции с этими понятиями. Глава 2 посвящена основам теории вероятностей и комбинаторики. В главе 3 излагаются основные понятия математической статистики, а также способы представления статистических данных. В главе 4 описываются характеристики случайных величин (математическое ожидание, среднеквадратическое отклонение, дисперсия, меры положения и меры рассеяния, коэффициент корреляции, моменты высших порядков) и математические операции над ними. В главе 5 представлены основные законы распределения случайной величины; особое внимание уделяется нормальному распределению и связанным с ним распределениям. В главе 6 вводится понятие точечной и интервальной оценки, в том числе, для угловых случайных величин. Глава 7 посвящена теории переноса ошибок. Глава 8 содержит краткую информацию из курса линейной алгебры (основные понятия и методы теории матриц). В главе 9 даются понятия «равноточные измерения», «неравноточные измерения», «условные уравнения» «нормальные уравнения», а также подробно разбирается общий метод линеаризации условных уравнений, решение соответствующих нормальных уравнений и представление решения. В главах 10–12 излагаются основы дисперсионного, корреляционного и регрессионного анализа. В том числе, излагается теория ортогональных

полиномов и поясняются преимущества их использования в полиномиальной регрессии. Глава 13 посвящена исследованиям вида распределения: критерию хи-квадрат (χ^2) и критерию Колмогорова. В последней главе 14 на примере прикладной задачи из области космологии (поиск космических струн) подробно рассмотрено применение непараметрических критериев. В приложении даны некоторые математические аспекты понятия фактора Байеса.

Пособие снабжено примерами преимущественно из астрономии, содержит выводы формул и доказательства теорем, которые обычно не приводятся в аналогичной литературе, но необходимы для более глубокого понимания математического смысла и обоснованности применения представленных статистических методов. Пособие является самодостаточным для решения широкого круга прикладных статистических задач без обращения к дополнительной литературе.

Помимо использованной в учебном пособии литературы [1]–[14], рекомендуется следующая дополнительная литература: по теории вероятностей и математической статистике [15]–[25], по линейной алгебе [26]–[28], по численным методам [29],[30].

Автор выражает большую благодарность проф. М.В. Сажину(ГАИШ МГУ им. М.В. Ломоносова) и проф. В.Е. Жарову (ГАИШ МГУ им. М.В. Ломоносова) за полезные обсуждения в процессе работы над рукописью, А.В. Моргуновой за помощь в работе, рецензентам Е.В. Шимановской и Е.А. Михайлову, а также студентам 1-го курса астрономического отделения физического факультета МГУ им. М.В. Ломоносова за внимательную вычитку текста.

1 Погрешности и ошибки измерений

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

Окружающий мир полон информации всевозможного рода. Качество сбора и обработки информации зависит не только от точности приборов и от надежности экспериментальных установок, но и от понимания того, что вся информация содержит элементы случайности. Невозможно провести несколько раз абсолютно одинаковые эксперименты или осуществить абсолютно одинаковые наблюдения и получить абсолютно одинаковые результаты. Случайность – это неотъемлемое свойство природы и избавиться от нее невозможно, поэтому необходимо уметь ее обнаруживать и контролировать, как качественно, так и количественно.

При наблюдениях, измерениях, экспериментах различают несколько видов ошибок.

1. *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.

2. *Субъективные ошибки* наблюдателя и экспериментатора, в том числе грубые ошибки и опечатки.

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

4. *Случайные ошибки*, которые могут быть как свойствами прибора, так и свойствами самого исследуемого процесса[1]. Эти ошибки изучаются статистическими методами.

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 повышением точности. Рассмотрим сначала, как оперировать результатами, лишенными случайной составляющей. Предположим, что существует точное числовое значение измеряемой величины (как не зависящая от прибора объективная

[1] Следует особо отметить важность случайных ошибок в современных исследованиях по до сих пор нерешенной проблеме возникновения жизни. Усложнение живых организмов также есть во многом случайный процесс. Ненаправленные случайные изменения (мутации) — это главный процесс, обеспечивающий материал для эволюции. Кроме того, на понятии случайности построен космологический антропный принцип и современные теории многомерных пространств (концепция Мультимира).

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

Предположим, выполнен 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), называется *прямой задачей*. 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

$$a - \varepsilon_a \leqslant A \leqslant a + \varepsilon_a,$$

называется *предельной абсолютной погрешностью*.

ПРИМЕР. Расстояние S от Земли до планеты Глизе 581c равно $a = 6{,}2$ парсека (1 пк $= 3 \cdot 10^{13}$ км)[2]. Округлим

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

до целого числа парсеков: $S \approx 6$ пк. Принято считать, что предельная абсолютная погрешность (ε_a) округленного приближенного значения равна половине единицы последнего знака округления, т.е. $\varepsilon_a = 1$ 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 $\varepsilon_a = 1$ км это очень хорошо, но та же величина для измерения длины Керченского моста представляет собой очень грубую погрешность). Введем понятие *предельной относительной погрешности*:

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

или

$$a(1 - \delta_a) \leqslant A \leqslant a(1 + \delta_a), \quad a > 0;$$

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

ПРИМЕР. 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%.

1.2.4 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:

$$\varepsilon_a = \sum_{i=1}^{n} \varepsilon_i.$$

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 $\varepsilon_a = \varepsilon_1 + \varepsilon_2$, а предельная относительная погрешность, соответственно,

$$\delta_a = \frac{\varepsilon_a}{|a|} = \frac{\varepsilon_1 + \varepsilon_2}{|a_1 - a_2|}. \tag{2}$$

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, например, умножением числителя и знаменателя (2) на выражение, сопряженное знаменателю.

ПРИМЕР. Пусть необходимо вычислить предельную относительную погрешность левой части формулы:

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

Это формула Эйлера [12], выражающая связь между двумя радиус-векторами параболы (r_1 и r_2), длиной хорды (s) и временем (u). Обычно s — малая величина по

сравнению с $r_1 + r_2$, поэтому прямое вычисление по формуле (2) приводит к потере точности. От разности близких чисел можно избавиться, умножив и разделив левую часть уравнения (3) на одну и ту же сумму

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

Тогда получаем

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

где

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

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

ПРИМЕР. 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. Последнее условие означает существование в каждой 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:

$$S = \iint\limits_{(D)} \sqrt{1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}\, dxdy. \qquad (4)$$

Действительно, угол γ между перпендикуляром и осью Oz определяется выражением

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

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

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

что дает формулу (4).

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

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

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

Кроме того,

$$\frac{\partial z}{\partial x} = -\frac{\partial F}{\partial x}\Big/\frac{\partial F}{\partial z};$$

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

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

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

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

Вводя полярные координаты на сфере единичного радиуса

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

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 $\alpha \in (0, 2\pi)$ 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

$$\alpha_1 = \alpha_3; \;\; \alpha_2 = \alpha_4; \;\; \delta_1 = \delta_2; \;\; \delta_3 = \delta_4.$$

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

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

$$\approx (\alpha_2 - \alpha_1) \cdot (\delta_2 - \delta_1) \cdot \cos\frac{\delta_1 + \delta_2}{2}.$$

Здесь и далее используем тригонометрическую формулу для устранения разности синусов близких углов. Поскольку следует 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 $\alpha_1^1, \alpha_2^1, \alpha_1^2, \alpha_2^2$ и $\delta_1^1, \delta_2^1, \delta_1^2, \delta_2^2$. Получаем два упорядоченных (*вариационных*) ряда

$$\tilde{\alpha}_1 < \tilde{\alpha}_2 < \tilde{\alpha}_3 < \tilde{\alpha}_4; \;\; \tilde{\tilde{\delta}}_1 < \tilde{\tilde{\delta}}_2 < \tilde{\tilde{\delta}}_3 < \tilde{\tilde{\delta}}_4.$$

Далее,

$$S_U = (\,\tilde{\alpha}_3 - \tilde{\alpha}_2\,) \cdot (\,\sin\tilde{\delta}_3 - \sin\tilde{\delta}_2\,) \approx$$

$$\approx (\,\tilde{\alpha}_3 - \tilde{\alpha}_2\,) \cdot (\,\tilde{\delta}_3 - \tilde{\delta}_2\,) \cdot \cos\frac{\tilde{\delta}_3 + \tilde{\delta}_2}{2}.$$

Рассмотрим два поля с координатами:

$$\alpha_1^1 = 170{,}56404;\ \ \delta_1^1 = 18{,}3216;$$
$$\alpha_2^1 = 169{,}51245;\ \ \delta_2^1 = 18{,}32824;$$
$$\alpha_3^1 = 170{,}5539;\ \ \delta_3^1 = 17{,}32329;$$
$$\alpha_4^1 = 169{,}50818;\ \ \delta_4^1 = 17{,}3299.$$

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 и 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

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

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

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 величина второго порядка малости, поэтому ей можно пренебречь. Окончательно получаем

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

или для предельной абсолютной погрешности

$$\varepsilon_a = |a_1| \varepsilon_2 + |a_2| \varepsilon_1.$$

Предельная относительная погрешность (1) произведения двух приближенных величин есть

$$\delta_a = \frac{\varepsilon_a}{|a|} = \frac{|a_1| \varepsilon_2 + |a_2| \varepsilon_1}{|a_1 a_2|} = \frac{\varepsilon_2}{|a_2|} + \frac{\varepsilon_1}{|a_1|} = \delta_1 + \delta_2.$$

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. Как и в случае вывода 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

$$\varepsilon_a = \frac{|a_1|\varepsilon_2 + |a_2|\varepsilon_1}{a_2^2}.$$

Предельная относительная погрешность (1) отношения двух величин есть

$$\delta_a = \left(\frac{|a_1|\varepsilon_2 + |a_2|\varepsilon_1}{a_2^2}\right) \cdot \left|\frac{a_2}{a_1}\right| = \frac{\varepsilon_1}{|a_1|} + \frac{\varepsilon_2}{|a_2|} = \delta_1 + \delta_2,$$

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 — некоторое число, такое что $a - \Delta_a \leqslant \xi \leqslant a$ (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. $\varepsilon_a \geqslant |\Delta_a|$, предельная абсолютная погрешность искомой функции

$$\varepsilon_u = |f'(a)|\varepsilon_a,$$

а предельная относительная погрешность искомой функции

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

Далее, предельная абсолютная погрешность функции нескольких аргументов имеет аналогичную структуру:

$$\varepsilon_u = \left|\frac{\partial f}{\partial x}\right|\varepsilon_a + \left|\frac{\partial f}{\partial y}\right|\varepsilon_b, \tag{5}$$

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

ПРИМЕР. Ускорение силы тяжести определяется с помощью оборотного маятника следующим образом:

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

где l — приведенная длина маятника, P — период колебания [12].

Наблюдения дали значения величин и их предельные абсолютные погрешности: $l = 50{,}02$ см; $P = 1{,}4196$ с; $\varepsilon_l = 10^{-2}$ см; $\varepsilon_P = 10^{-4}$ с.

Вычислим ускорение силы тяжести g и его предельную абсолютную погрешность (значение $\pi = 3{,}1416$, т.е. $\varepsilon_\pi = 5 \cdot 10^{-5}$).

Применяя формулу (5), получим выражение для предельной абсолютной погрешности:

$$\varepsilon_g = \frac{8\pi l}{P^2}\varepsilon_\pi + \frac{4\pi^2}{P^2}\varepsilon_l + \frac{8\pi^2 l}{P^3}\varepsilon_P = 0{,}37 \text{ см/с}^2.$$

После выполнения вычислений получим

$$g = 979{,}88 \pm 0{,}37 \,\text{см/с}^2.$$

Если погрешность вычислена с точностью до одной значащей цифры, $\varepsilon_g = 0{,}4$ см/с^2, то

$$g = 979{,}9 \pm 0{,}4 \,\text{см/с}^2.$$

2 Основы теории вероятности и комбинаторики

2.1 Опыт, событие и вероятность

Основными объектами, 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 опыта событие появляется с определенной частотой. Частота служит для определения основного понятия теории вероятности — собственно *вероятности события*. Пусть проведено n одинаковых опытов («одинаковость» означает одинаковые условия всех опытов). Пусть результатами этих опытов служит 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 называется вероятностью события[3] A:

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

Внимательный читатель заметит, что граничные значения для p есть 0 и 1.

Если событие невозможно, то его вероятность равна 0, однако событие с вероятностью $p = 0$ необязательно

[3] Строгое определение вероятности см., например, в [9].

невозможно: так, вероятность попасть, стреляя из пистолета, в заданную точку на стене (без какой-либо погрешности) равна нулю, но это событие не является невозможным. Невозможное событие обозначается $\varnothing$. Аналогично вероятность достоверного события равна 1, достоверное событие 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, с помощью аппарата теории множеств [9]. Комбинаторика и теория вероятностей представляют собой обширные самостоятельные дисциплины, поэтому здесь будут введены только основные понятия, а также будут обсуждаться 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 подобных 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; G — область всех возможных исходов.

ПРИМЕР. В любые моменты промежутка времени от 0 до T равновозможны поступления в приемник двух сигналов [1]. Приемник будет забит, если промежуток

времени между моментами поступления сигналов меньше τ. Нужно 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, следовательно, область их допустимых значений представляет собой квадрат со стороной T. Область, соответствующая тому, что приемник забит (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:

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

Очевидно, если $\tau \to T$, то вероятность p стремиться к единице; если $\tau \to 0$, то вероятность p стремиться к нулю.

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

Вероятность произведения событий означает вероятность того, что произойдет и одно событие, и другое. Произведение событий обозначается $A \cdot B$ или $A \cap B$.

Используя понятие условной вероятности, рассмотрим операцию *умножения вероятностей*:

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

Если A и B — независимы, то

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

2.3.3 Сложение вероятностей

Вероятность суммы событий означает вероятность того, что произойдет или одно событие, или другое. Сумма событий обозначается $A + B$ или $A \cup B$.

Если события A_i и A_j *несовместные*, т.е.

$$A_i \cdot A_j = \varnothing \ \ (i \neq j),$$

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. Так, если A_i и A_j *независимые*, 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 $1 - P(A + B) = P(\bar{A} \cdot \bar{B})$, где чертой обозначено *противоположное событие* (дополнительное событие), т.е. $\bar{A} = \Omega - A$. Действительно,

$$1 - P(A + B) = 1 - P(A) - P(B) + P(A \cdot B).$$

С другой стороны, т.к. события A и B независимы, то противоположные им события $\bar{A}$ и $\bar{B}$ также независимы и для них выполняется

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

Раскрывая скобки в последнем выражении, получаем искомое равенство.

ПРИМЕР. По многолетним наблюдениям известна вероятность того, что в районе обсерватории ночь будет ясной [1]: 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:

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

Для решения задача сначала сформулируем происходящие события:

- A_1 — ясная ночь с 5 на 6 февраля;
- A_2 — ясная ночь с 20 на 21 февраля;
- B_1 — ясная ночь с 5 на 6 марта;
- B_2 — ясная ночь с 20 на 21 марта;
- C_1 — ясная ночь с 5 на 6 апреля;
- C_2 — ясная ночь с 20 на 21 апреля.

Поскольку предоставленные астроному ночи для наблюдения отделены друг от друга большим периодом времени (15 дней), то можно рассматривать события (наблюдения) независимыми. Фраза о том, что будет одна ясная ночь, эквивалентна тому, что ясной будет *хотя бы одна* ночь. Следовательно, вероятность того, что будет одна ясная ночь есть

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

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

$$\begin{aligned} p_1 &= P(A_1 + A_2 + B_1 + B_2 + C_1 + C_2) = \\ &= 1 - P(\bar{A}_1)P(\bar{A}_2)P(\bar{B}_1)P(\bar{B}_2)P(\bar{C}_1)P(\bar{C}_2) = \\ &= 1 - (1 - 0{,}18) \cdot (1 - 0{,}24) \cdot (1 - 0{,}36) \cdot (1 - 0{,}18) \times \\ &\times \ (1 - 0{,}24) \cdot (1 - 0{,}36) \approx 0{,}84. \end{aligned}$$

Здесь было учтено, что все события независимы, а вероятность произведения независимых событий равна произведению вероятностей.

Вероятность p_2 того, что будут две ясные ночи, очевидно, равна

$$1 - P(\text{ни одна ночь не ясная}) - P(\text{ровно одна ночь ясная}).$$

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 Оценка на вероятность произведения событий

Следующее полезное соотношение представляет собой оценку снизу на вероятность произведения двух событий:

$$P(A \cdot B) \geqslant P(A) + P(B) - 1.$$

Для доказательства этого утверждения заметим, что

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

Следовательно, т.к. $P(A+B) \leqslant 1$,

$$P(A \cdot B) = P(A) + P(B) - P(A+B) \geqslant P(A) + P(B) - 1.$$

2.5 Полная вероятность

Следствием правил сложения и умножения вероятностей является правило *полной вероятности*. 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 (гипотез): $H_1, H_2, H_3, \ldots, H_n$. Интересно, что гипотезы могут быть сформулированы разными способами, важно помнить, что они должны быть взаимоисключающими (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, как заметит внимательный читатель, означает

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

или

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

Каждая гипотеза H_i — это случайное событие, вероятность которого до проведения опыта (*априорная вероятность*) оценивается как $P(H_i)$.

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

Тогда

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

потому что (в силу того, что гипотезы по определению полностью охватывают все возможные условия опыта) событие A может появиться только вместе с одной из гипотез

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

Кроме того,

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

Таким образом, формула *полной вероятности*:

$$P(A) = \sum_{i=1}^{n} P(H_i) \cdot P(A|H_i). \tag{6}$$

ПРИМЕР. Среди наблюдаемых спиральных галактик 23% принадлежат подтипу Sa; 31% — подтипу Sb; 46% — подтипу Sc [1]. Вероятность вспышки в течение года сверхновой звезды в галактике Sa составляет 0,0020; в галактике Sb — 0,0035; в галактике Sc — 0,0055. Нужно 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 (6)

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

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:

$$P = 0{,}23 \cdot 0{,}0020 + 0{,}31 \cdot 0{,}0035 + 0{,}46 \cdot 0{,}0055 = 0{,}0041.$$

2.6 Формула Байеса

Рассмотрим *формулу Байеса* как следствие формулы полной вероятности и формулы умножения вероятностей. 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 $P(H_k|A)$ — *апостериорную вероятность*.

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

Пусть событие A произошло. Тогда можно заново пересчитать вероятности $P(H_1), P(H_2), \ldots, P(H_n)$ с учетом того, что событие A произошло.

Найдем $P(H_1|A)$, $P(H_2|A)$, $\ldots$, $P(H_n|A)$. 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. Дополним предыдущую задачу о галактиках новой информацией: пусть в далекой спиральной галактике была обнаружена вспышка сверхновой. Требуется найти вероятности того, что галактика принадлежит одному из подтипов S_a, S_b или S_c.

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

$$P(S_a|H) = \frac{0{,}23 \cdot 0{,}0020}{0{,}0041} = 0{,}11;$$

$$P(S_b|H) = \frac{0{,}31 \cdot 0{,}0035}{0{,}0041} = 0{,}26;$$

$$P(S_c|H) = \frac{0{,}46 \cdot 0{,}0055}{0{,}0041} = 0{,}62.$$

2.7 Элементы комбинаторики

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

В комбинаторике существуют две принципиально различные схемы такого выбора. В первой схеме каждый выбранный элемент исключается из исходного множества. Во второй схеме происходит поэлементный выбор с обязательным возвращением выбранного элемента на каждом шаге и перемешиванием. После осуществления выбора согласно одной из двух схем, элементы могут быть упорядочены или не упорядочены.

Таблица 1

Основные формулы комбинаторики.

	Размещения	Перестановки	Сочетания
без повторений	$A_n^k = \frac{n!}{(n-k)!}$	$P_n = n!$	$C_n^k = \frac{n!}{k!(n-k)!}$
с повторениями	$\bar{A}_n^k = n^k$	$P(n_1, n_2, \ldots, n_k) = = \frac{n!}{n_1! n_2! \ldots n_k!}$	$\bar{C}_n^m = \frac{(n+m-1)!}{(m-1)!n!}$

Более подробно рассмотрим вышесказанное на примерах.

Так, **размещения без повторений**[4] 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**[5], когда повторение цифр разрешено, получаем $\bar{A}_5^4 = 5^4$. Приведем пример из статистической физики [7]. 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 классической статистики Максвелла–Больцмана, в которой частицы различимы, каждая из частиц может попасть в любую из m ячеек независимо от остальных частиц. Тогда число всех возможных со-

[4] Другие названия: упорядоченная выборка без повторений, упорядоченная выборка без возвращений.

[5] Другие названия: упорядоченная выборка с повторениями, упорядоченная выборка с возвращениями.

стояний такой системы есть $\bar{A}_m^n = m^n$.

Задачу на **перестановки без повторений** можно рассматривать как частный случай задачи на размещения без повторений, 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 четыре 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 без повторений[6] — это наиболее часто используемая в статистике комбинаторная формула (схема испытаний Бернулли). Расчет числа сочетаний без повторений основан на использовании коэффициентов разложения бинома Ньютона:

$$(a+b)^n = \sum_{k=0}^{n} C_n^k \cdot a^{n-k} \cdot b^k; \;\; C_n^k = \frac{n!}{k!(n-k)!}$$

Биномиальные коэффициенты C_n^k обладают очевидными свойствами:

$$C_n^k = C_n^{n-k}; \;\; C_{n+1}^{k+1} = C_n^{k+1} + C_n^k,$$

[6] Другие названия: неупорядоченная выборка без повторений, неупорядоченная выборка без возвращений.

а соответствующие факториалы при больших n вычисляются приближенно по *формуле Стирлинга*:

$$n! \to \sqrt{2\pi n} \cdot n^n \cdot e^{-n}.$$

Очевидно, что количество сочетаний без повторений меньше, чем количество размещений без повторений, т.к. в последних важен еще и порядок. Типичной задачей на применение этой формулы является выбор шаров из урны. Например, из урны, содержащей $n = 10$ шаров, можно наугад выбрать $k = 4$ шара, при этом количество способов есть $C_n^k = C_{10}^4 = 10!/4!/6! = 210$. Снова используя пример из статистической физики, рассмотрим квантовую статистику Ферми–Дирака (справедлива для электронов, нейтронов, протонов), при которой n частиц неразличимы, их число меньше числа ячеек ($n < m$) и каждая ячейка может содержать не более одной частицы. Тогда такая квантово-механическая система характеризуется числом равновозможных состояний C_m^n.

В заключение рассмотрим несколько примеров задач на **сочетания с повторениями**[7]. Число таких сочетаний — это количество наборов элементов, в которые каждый элемент может входить несколько раз. 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, $\{1\,1\,0\,1\,1\,1\,1\,0\,1\,1\,0\,1\,1\}$ означает, что было набрано $k_1 = 2$ деталей первого типа, $k_2 = 4$ деталей второго типа,

[7] Другие названия: неупорядоченная выборка с повторениями, неупорядоченная выборка с возвращениями.

$k_3 = 2$ 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 десяти единиц и трех нулей: $\bar{C}_{10}^4 = P(10, 3) = 13!/3!/10! = 286$.

Возвращаясь к примеру из статистической физики, рассмотрим квантовую статистику Бозе–Эйнштейна (справедлива для фотонов, атомных ядер, атомов с четным числом элементарных частиц), в соответствии с которой все n частиц неразличимы и все их распределения по m ячейкам равновозможны. Такая квантово-механическая система характеризуется числом состояний C_{m+n-1}^n.

Еще один пример: уравнение $x_1 + x_2 + \cdots + x_m = n$ при натуральном n имеет C_{m+n-1}^n неотрицательных целочисленных решений[8].

ПРИМЕР. Приведем простой 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 [1]. Нужно определить вероятность того, что из двух объектов, наугад выбранных в каталоге,

1. каждый окажется галактикой;

2. один окажется шаровым, а другой — рассеянным скоплением.

В первом случае вероятность определяется отношением числа способов выбрать два объекта из имеющихся

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

39 галактик (*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 шаровых скоплений и одновременный выбор одного объекта из 29 рассеянных скоплений. Общее возможное количество вариантов выбрать два объекта из всего каталога определяется так же, как в предыдущем пункте:

$$p_b = \frac{C_{29}^1 C_{29}^1}{C_{108}^2} \approx 0{,}146.$$

ПРИМЕР. В заключение этого раздела рассмотрим использование идеи *симметрии* в теории вероятностей.

Из множества $\{1, 2, ..., 100\}$ последовательно и без возвращений выбирают два числа. Какова вероятность, что второе число окажется больше первого?

Исходы опыта — это упорядоченные пары (a, b) неравных друг другу чисел из множества $\{1, 2, ..., 100\}$. Соответствие $(a, b) \to (b, a)$ показывает, что количество пар с первой большей составляющей такое же, как и количество пар с большей второй составляющей, т.е. искомая вероятность есть 0,5.

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* типа.

Обратим внимание на обозначения: прописными латинскими буквами

$$X, Y, Z, \ldots$$

будем обозначать сами случайные величины, а строчными латинскими буквами

$$x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n, z_1, z_2, \ldots, z_n$$

— возможные значения, которые могут принимать эти случайные величины или, что то же самое, результаты эксперимента или наблюдения некоторой случайной величины.

ПРИМЕР. Число очков, выпавших при однократном бросании кубика, есть случайная величина дискретного типа, множество возможных значений которой: $\{x_1, x_2, x_3, x_4, x_5, x_6\} = \{1, 2, 3, 4, 5, 6\}$.

ПРИМЕР. Ошибка измерения скорости кометы Δv есть случайная величина непрерывного типа, множество возможных значений которой принадлежит интервалу $[\Delta v_{\min}, \Delta v_{\max}]$.

3.1.2 Генеральная совокупность

Генеральной совокупностью называется набор всех возможных значений случайной величины (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.

Выборка, упорядоченная по какому-либо принципу, называется *вариационным рядом*.

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.

Для дискретного распределения, каждому значению случайной величины $\{x_1, x_2, \ldots, x_n\}$ ставится в соответствие своя вероятность $\{p(x_1), p(x_2), \ldots, p(x_n)\}$, 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.

1. *ряд распределения* $\{x_i, p(x_i)\}$;
2. *функция распределения* $F(x)$, или интегральный закон распределения;

Таблица 2

Представление закона распределения случайной величины в виде таблицы — статистического ряда распределения.

x_i	x_1	x_2	x_3	...	x_n
$p(x_i)$	$p(x_1)$	$p(x_2)$	$p(x_3)$	...	$p(x_n)$

3. *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 неслучайного аргумента (т.е. сами не являются случайными). Они есть функции значений, которые может принимать случайный аргумент.

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* (см. табл. 2). Каждому значению x_i ставится в соответствие вероятность $p(x_i)$.

ПРИМЕР. Пусть производится два независимых опыта, в каждом из которых событие A появляется с вероятностью $p = 0{,}60$. Нужно построить закон распределения случайной величины X — числа появлений события A — в виде ряда распределения.

Исходя из условия задачи, случайная величина X может принимать значения $\{x_0, x_1, x_2\} = \{0, 1, 2\}$. Найдем

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

соответствующие вероятности $\{p(x_0), p(x_1), p(x_2)\}$:

1. $p(x_0)$ есть вероятность того, что ни в первом, ни во втором случае событие A не появилось;

2. $p(x_1)$ есть вероятность того, что событие A появилось ровно один раз (либо в первом опыте, либо во втором);

3. $p(x_2)$ 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, (см. табл. 3).

Таблица 4

Пример схемы распределения (1).

x_i	x_1	x_2
$p(x_i)$	0,5	0,5

3.1.6 Энтропия конечной схемы

Для реальной выборки (или конечной схемы), когда число всех возможных значений, принимаемых случайной величиной, конечно и равно n, можно выписать соответствующую таблицу данных. Если к тому же все значения случайной величины равноточны (имеют одинаковую точность), то они представимы в виде табл. 4 (простой пример для двух значений).

Можно определить *энтропию конечной схемы*. Всякая конечная схема описывает некое состояние неопределенности в том смысле, что нам известны только вероятности возможных значений случайной величины. В разных схемах степень этой неопределенности различна. Например, схема из табл. 4 более неопределенна, чем схема из табл. 5.

Удобной мерой степени неопределенности служит величина энтропии конечной схемы:

$$E = -\sum_{k=1}^{n} p(x_k) \cdot \lg p(x_k).$$

Полагается, что если $p(x_k) = 0$, то $p(x_k) \cdot \lg p(x_k) = 0$. Если из $p(x_i)$ какое-то значение равно 1, а все остальные равны 0, то энтропия есть ноль, т.е. неопределенность отсутствует. Неопределенность максимальна, когда все $p(x_i) = 1/n$ $(i = 1, 2, ...n)$.

Таблица 5

Пример схемы распределения (2).

x_i	x_1	x_2
$p(x_i)$	0,98	0,02

3.1.7 Функция распределения

Функцией распределения случайной величины X называется функция $F(x)$, определенная на всей действительной оси следующим образом:

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

где X — случайная величина; 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

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

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

1. вероятность невозможного события равна нулю:

$$F(-\infty) = P(X < -\infty) = P(\varnothing) = 0;$$

2. вероятность достоверного события равна единице:

$$F(+\infty) = P(X < +\infty) = P(\Omega) = 1;$$

3. функция распределения неубывающая: для $x_2 > x_1$ выполняется $F(x_2) \geqslant F(x_1)$.

Дискретный аналог функции распределения (понятие функции распределения для дискретных величин) — это *кумулята*.

ПРИМЕР. Построим функцию распределения для предыдущего примера:

$$F(0) = P(X < 0) = 0;$$

$$F(1) = P(0 \leqslant X < 1) = P(X = 0) = 0{,}16;$$

$$F(2) = P(0 \leqslant X < 2) = P(X = 0) + P(X = 1) =$$

$$= 0{,}16 + 0{,}48 = 0{,}64;$$

$$F(2 + \varepsilon) = P(0 \leqslant X < 2 + \varepsilon) =$$

$$= P(X = 0) + P(X = 1) + P(X = 2) =$$

$$= 0{,}16 + 0{,}48 + 0{,}36 = 1.$$

Функция F определена на всей действительной оси.

3.1.8 Плотность вероятности

Плотность вероятности (также *плотность распределения*, *функция плотности 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:

1. $f(x) \geqslant 0$;
2. $\int_{-\infty}^{+\infty} f(x)dx = 1$.

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

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

Дискретный аналог плотности вероятности — *гистограмма.*

3.1.9 Двумерное распределение

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

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

Двумерная функция распределения есть

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

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

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

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 кусочно-заданную функцию

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

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

$$\int_0^1\int_0^1 \big(x+y\big)dxdy=\int_0^1\left(\int_0^1 xdx\right)dy+\int_0^1\left(\int_0^1 ydx\right)dy=$$

$$=\int_0^1\frac{1}{2}dy+\int_0^1 ydy=\frac{1}{2}+\frac{1}{2}=1.$$

ПРИМЕР. Пусть плотность распределения имеет вид [14]:

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

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,

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

$$=c\int_{-1}^1 x^2\left(\int_{x^2}^1 ydy\right)dx=c\int_{-1}^1 x^2\frac{1-x^4}{2}dx=\frac{4c}{21}.$$

Следовательно,

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

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

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

$$= \frac{21}{4}\int_0^1 x^2 \left(\frac{x^2 - x^4}{2}\right) dx = \frac{3}{20}.$$

ПРИМЕР. Рассмотрим вычисление одномерной плотности по известной двумерной плотности. Пусть

$$f(x,y) = e^{-x-y};\ \ x, y \geqslant 0.$$

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 [14]:

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

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

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

Тогда

$$P(X + Y \leqslant 1) = \iint\limits_{x+y \leqslant 1} f(x,y) dx dy =$$

$$4\int_0^1 x \left(\int_0^{1-x} y dy\right) dx = 4\int_0^1 x \frac{(1-x)^2}{2} dx = \frac{1}{6}.$$

ПРИМЕР. Пусть двумерная плотность задана в виде [14]:

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

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

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

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 (см. табл. 6) как соответствие номера наблюдения i и результата наблюдения x_i.

3.2.2 Вариационный ряд

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

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

то полученный ряд будет называться *вариационным рядом*.

Величина x_k^* называется *«порядковая статистика»*. Величина $J_n(x) = x_n^* - x_1^*$ называется *«размах выборки»*. В дальнейшем *«статистика»* — любая величина, имеющая заданный закон распределения.

Таблица 6

Представление простого статистического ряда.

Номер наблюдения i	1	2	...	n
Результат наблюдения x_i	x_1	x_2	...	x_n

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; n — объем выборки (общее количество элементов выборки).

Величина X может принимать и одинаковые значения. Тогда пусть k — число разных значений величины X (очевидно, $k \leqslant n$). Пусть индекс $\nu = \{1, 2, \ldots, k\}$. 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. Пусть для i, принимающих значения

$$\{1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20\},$$

соответствующие $\{x_i\}$ есть

$$\{1; -2; 0; 1; -3; -1; 0; -1; 1; 3; 0; -1; 1; 2; 0; -1; 0; -2; 0; 1\}.$$

Случайная величина принимает всего семь различных значений с определенной частотой, что должно быть отражено в виде табл. 7 для построения эмпирической функции распределения.

Таблица 7

Таблица для построения эмпирической функции распределения.

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

3.2.4 Полигон частот. Алгоритм построения полигона частот

Полигон частот – это сгруппированные данные выборки. Если объем выборки $\{x_1, x_2, x_3, \ldots, x_n\}$ 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.

Алгоритм построения полигона частот, из которого станет ясно его точное определение, выглядит следующим образом.

1. Построить вариационный ряд данных (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 группировки. Число интервалов можно выбрать: $k \approx \log_2 n + 1$. В практических задачах $7 \leqslant k \leqslant 10$. Иногда удобно взять интервалы разной длины в зависимости от количества попадающих в них точек.

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

4. Подсчитать количества выборочных данных, попавших в каждый интервал (см. табл. 8): $n_1, n_2, \ldots, n_k$.

Таблица 8

Количество попаданий значений случайной величины в построенные интервалы.

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

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

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

3.2.5 Гистограмма

Гистограмма — это дискретный аналог[9] функции плотности вероятности, называемая также эмпирической плотностью вероятности $f^*(x)$.

Пусть p_j^* есть площадь прямоугольника с длиной основания $\Delta\hat{x}_j = \hat{x}_j - \hat{x}_{j-1}$ $(j = 1, \ldots, k)$. Тогда высота этого прямоугольника — 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 частот, точки $\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_k$ — крайние точки интервалов, на которые разбивается вариационный ряд обрабатываемых данных. Совокупность таких прямоугольников для всех $\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_k$ составляет гистограмму, которая при большом количестве точек переходит в свой непрерывный

[9] 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.

аналог — в функцию плотности вероятности. Сумма площадей всех прямоугольников равна единице.

3.2.6 Кумулята

Кумулята — приближенная эмпирическая функция распределения. Как гистограмма является дискретным аналогом функции плотности вероятности, так и кумулята является дискретным аналогом функции распределения:

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

где $\hat{x}_0, \hat{x}_1, \ldots, \hat{x}_\nu, \ldots, \hat{x}_k$ — крайние точки интервалов, на которые разбивается вариационный ряд обрабатываемых данных; $k \approx \log_2 n + 1$ — стандартно рекомендуемое число интервалов разбиения; n — число элементов выборки; n_j — число данных измерений, попавших в интервал $[\hat{x}_{j-1}, \hat{x}_j]$. Точка $\hat{x}_\nu$ — один из концов интервала: при вычислении кумуляты все предыдущие вероятности, как заметит внимательный читатель, складываются. То же самое происходит 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* [14], заключающегося в минимизации оценки скользящего контроля J:

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

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

где $f^*_{(-i)}(\Delta\hat{x}_j)$ — гистограмма, построенная после удаления i-го наблюдения из массива данных.

Минимизируемая функция требует пересчета гистограммы n раз, что представляет собой довольно громоздкую процедуру численного счета. Для упрощения расчетов оценка скользящего контроля может быть представлена в виде [14]:

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

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

3.2.8 Ядерная оценка плотности

В отличие от классической гистограммы, метод *ядерной оценки плотности* [14] представляет собой сглаженную оценку плотности распределения:

$$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\cdot K\Big(\frac{x-x_i}{h}\Big)\ \ (h>0).$$

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

1. $K \geqslant 0$;

2. $\int xK(x)dx = 0$;

3. $\sigma_K^2 \equiv \int x^2 K(x)dx > 0$.

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

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

где

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

ПРИМЕР. Ядро Епанечникова.

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

ПРИМЕР. Гауссово ядро.

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

В этом случае величина

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

т.е. имеет нормальный закон распределения (см. раздел 5.14) со средним 0 и дисперсией 2.

4 Характеристики случайных величин

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, μ_x, μ 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 $\{x_1, x_2, \ldots, x_n\}$ с соответствующими вероятностями $\{p(x_1), p(x_2), \ldots, p(x_n)\}$, 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 все остальные характеристики случайной величины, как будет показано ниже.

ПРИМЕР. Не каждое распределение обладает конечным математическим ожиданием. Рассмотрим рас-

пределение Коши, плотность распределения которого задается функцией

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

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

$$M[X] = \frac{2}{\pi}\int_0^{+\infty}\frac{xdx}{1+x^2} = \frac{1}{\pi}\Big[\ln(x^2+1)\Big]\Big|_0^{\infty} = \infty.$$

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. На графике плотность распределения Коши обладает широкими «крыльями», не спадающими к нулю на бесконечности, а значит есть 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 обладает следующими свойствами:

1. $M[C] = C$;
2. $M[C \cdot X] = C \cdot M[X]$;
3. $M[X \pm Y] = M[X] \pm M[Y]$;
4. Пусть X, Y — непрерывные случайные величины и $Y = g(X)$. Тогда

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

5. 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:

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

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

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; условное математическое ожидание $M[X|Y = y]$ — это неслучайная функция переменной y.

Для функции случайных аргументов $g(X, Y)$ (для дискретного и непрерывного случая, соответственно)

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

$$= \sum_{i=1}^{n} g(x_i, y_i) \cdot p(x_i|y) = \sum_{i=1}^{n} g(x_i, y_i) \cdot \frac{p(x_i, y)}{p(y)};$$

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

4.2 Среднеквадратическое отклонение

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

В литературе среднеквадратическое отклонение (или стандартное отклонение) случайной величины X обозначается обычно $s.d.$, σ_x, σ, $\sigma[X]$, s.

По определению, среднеквадратическое отклонение дискретной случайной величины X, принимающей значения $\{x_1, x_2, \ldots, x_n\}$ с соответствующими вероятностями $\{p(x_1), p(x_2), \ldots, p(x_n)\}$, 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 обычно обозначается $D[X]$, σ^2 или σ_x^2, s^2.

По определению, дисперсия дискретной случайной величины X, принимающей значения $\{x_1, x_2, \ldots, x_n\}$ с соответствующими вероятностями $\{p(x_1), p(x_2), \ldots, p(x_n)\}$, 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 обладает следующими свойствами:

1. $D[C] = 0$;

2. $D[C \cdot X] = C^2 \cdot D[X]$;

3. $D[X] = M[X^2] - (M[X])^2$;

4. $D[X \pm Y] = D[X] + D[Y] \pm 2M[(X - m_x) \cdot (Y - m_y)]$.
   Величина $M[(X - m_x) \cdot (Y - m_y)]$ называется *ковариацией*. Ковариация равна нулю, если X и Y — независимые случайные величины.

5. $D[XY] = D[X] \cdot D[Y] + m_x^2 D[Y] + m_y^2 D[X]$, если X и Y — независимые случайные величины.

Докажем последнее свойство. По определению дисперсии, для случайной величины $Z = X \cdot Y$

$$D[Z] = M[(Z - m_z)^2] =$$

$$= M[(XY)^2 - 2XY \cdot M[XY] + (M[XY])^2],$$

где $m_z = M[Z] = M[XY]$ есть постоянная величина, математическое ожидание произведения двух случайных величин. Далее, используя свойство линейности математического ожидания и тот факт, что математическое ожидание постоянной величины есть сама эта величина, получаем

$$D[XY] = M[(XY)^2] - 2M[XY] \cdot M[XY] + (M[XY])^2.$$

Поскольку X и Y — независимые случайные величины, то $M[XY] = M[X]M[Y]$, следовательно,

$$D[XY] = M[(XY)^2] - 2M[XY] \cdot M[XY] + (M[XY])^2 =$$

$$= M[(XY)^2] - (M[X] \cdot M[Y])^2.$$

Далее распишем правую часть доказываемого соотношения:

$$D[X] \cdot D[Y] + m_x^2 D[Y] + m_y^2 D[X] =$$

$$= M[(X - m_x)^2] \cdot M[(Y - m_y)^2] + m_x^2 M[(Y - m_y)^2] +$$
$$+ m_y^2 M[(X - m_x)^2] = \Big(M[X^2] - (M[X])^2\Big) \cdot \Big(M[Y^2] -$$
$$- (M[Y])^2\Big) + m_x^2 M[Y^2] - m_x^2 (M[Y])^2 + m_y^2 M[X^2] -$$
$$- m_y^2 (M[X])^2 = M[X^2] \cdot M[Y^2] + (M[X] \cdot M[Y])^2 -$$
$$- 2m_x^2 m_y^2 = M[(XY)^2] - (M[X] \cdot M[Y])^2.$$

4.3.2 Условная дисперсия

Величина

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

где $\mu(x) = M[Y|X = x]$, называется *условной дисперсией*, [14].

4.4 Меры положения

Такие характеристики случайной величины как *среднее, взвешенное среднее, медиана* и *мода* характеризуют примерное положение истинного значения искомой величины X и поэтому носят общее название *меры положения*. Для некотрых распределений, например, для *нормальноro* (*гауссового*) распредления (см. раздел 5.14) все эти величины совпадают.

4.4.1 Среднее

Пусть некоторая величина X наблюдается или измеряется некоторым прибором n раз. При статистической обработке выборки $\{x_1, x_2, \ldots, x_n\}$, если не оговорено особо, все значения x_i считаются равноправными (равновероятными, *равноточными*). Таким образом, для оценки

среднего значения (обозначается $\bar{x}$) искомой величины X применяется формула для математического ожидания с учетом того, что $p(x_i) = p = 1/n$:

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

Последняя формула, очевидно, есть среднее арифметическое всех элементов выборки.

4.4.2 Взвешенное среднее

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

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

4.4.3 Медиана

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

1. $x_m = x_{(n+1)/2}$, если n — нечетное,
2. $x_m = \frac{1}{2}\big(x_{n/2} + x_{(n/2)+1}\big)$, если n — четное.

Число x_m называется *медианой*.

4.4.4 Мода

Среднее может быть оценено по наиболее часто встречающемуся элементу выборки, т.е. элементу x_l, при котором плотность вероятности ($f(x)$) максимальна. Такое x_l называется *мода*.

Заметим, что для исследования распределений, имеющих два четко выраженных максимума (*бимодальное* распределение) или более (*мультимодальное* распределение) мода — лучшая характеристика, чем математическое ожидание, поскольку последнее не даст информации об указанных пиках. В практических задачах для удобства дальнейшей обработки такие распределения могут быть представлены композицией нескольких распределений.

Кроме того, оценивание не среднего значения распределения, а его моды является важным методом получения устойчивых оценок параметров распределений при достаточно больших выборках. Это обусловлено тем, что в этом случае мода, т.е. положение максимума распределения, определяется более представительным количеством членов выборки с малыми отклонениями от моды, а крылья распределения (возможно, содержащие аномальные выбросы) не влияют на положение моды.

4.5 Меры рассеяния

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:

1. $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ — *среднеквадратическое отклонение*;

2. $d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$ — *среднее отклонение*;

3. $r = J_n(x) = x_n^* - x_1^* = x_{\max} - x_{\min}$ — *размах*, все x_i предполагаются с весом 1.

Отдельно 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; $\sigma_{\bar{x}}$ 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.

4.6 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

$$X = \{x_1, x_2, \ldots, x_n\}, \;\; Y = \{y_1, y_2, \ldots, y_n\}.$$

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

4.7 Моменты случайных величин

И математическое ожидание, и дисперсия случайной величины X представляют собой частные 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*:

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

Очевидно,

$$\alpha_1[X] = M[X] = \mu;$$
$$\mu_2[X] = D[X] = \sigma^2.$$

Моменты вычисляются по определению математического ожидания. Начальные моменты для дискретного распределения и для непрерывного распределения, соответственно:

$$\alpha_k[X] = \begin{cases} \sum\limits_{i=1}^{n} x_i^k \cdot p(x_i); \\ \int\limits_{-\infty}^{\infty} x^k f(x) dx. \end{cases}$$

Центральные моменты для дискретного распределения и для непрерывного распределения, соответственно:

$$\mu_k[X] = \begin{cases} \sum\limits_{i=1}^{n} (x_i - \mu)^k \cdot p(x_i); \\ \int\limits_{-\infty}^{\infty} (x - \mu)^k f(x) dx. \end{cases}$$

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

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

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

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

Для нормального распределения (см. раздел 5.14) $\gamma_1 = \gamma_2 = 0$. Отличие от нуля этих характеристик позволяет судить о том, насколько исследуемое распределение отличается от нормального распределения.

4.8 Распределение вероятности для функции случайных величин

4.8.1 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 [11].

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

Таблица 9

Статистический ряд случайной величины X.

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

табл. 9). Тогда закон распределения $h(X)$ будет определяться из следующих выражений:

$$h_i = h(x_i);$$

$$p_i^H = P\big(\, h(x_i) = h_i \,\big) = P(X = x_i).$$

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

$$p_1^H =$$

$$= P(h(x_1) = h_1) = P(h(0) = \cos 0 = 1) = P(X = 0) = \frac{1}{6};$$

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

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

Результат вычислений см. табл. 10, где h_i расположены в порядке возрастания.

Важно отметить, что приведенные рассуждения верны, если существует только одно значение $X = x_k$, при котором $h(X) = h(x_k) = h_0$. Если существует несколько

Таблица 10

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 11

Статистический ряд случайной величины Y.

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

значений $X = x_\nu, x_{\nu+1}, \ldots, x_{k-1}, x_k$, при которых $h(X) = h_0$ (функция, обратная h, не однозначная), то

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

Проиллюстрируем сказанное, расширив выборку (см. табл. 11 – 12) для дискретной случайной величины X из предыдущего примера и переобозначив X на Y. (Таблицы приведены без дополнительных пояснений в тексте.)

Таблица 12

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

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

4.8.2 Непрерывная случайная величина

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

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

Эта формула очевидна, потому что для непрерывной случайной величины плотность распределения $f(x)$ есть производная функции распределения $F(x)$; функция распределения в данном случае рассматривается на малом интервале $[x; x + dx]$.

Ставится задача [11] найти плотность распределения $g(h)$, такую, что

$$g(h)dh = P(h \leqslant H \leqslant h + dh); \;\; h = h(x).$$

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 $|\ldots|$ — 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 есть $f(x)dx = a+bx$ $(0 \leqslant x \leqslant \pi/2)$. Найдем плотность вероятности $g(h)$:

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

$$0 \leqslant h \leqslant 1.$$

Окончательно получаем

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

ПРИМЕР. Вероятность обнаружить звезду в объеме dv равна $k \cdot dv$. Для каждой звезды найдется другая звезда — ее ближайший сосед [1]. Требуется 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 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 на $k \cdot 4\pi \cdot x^2 dx$ обе части последнего уравнения, потом продифференцируем по x и получим

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

Здесь мы учли, что, по определению плотности распределения и функции распределения, $F'(x) = f(x)$.

После интегрирования обеих частей последнего равенства получим

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

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 из таблиц.

Дисперсия:

$$\sigma^2 = \int_0^\infty (x-\bar{x})^2 \cdot f(x)dx =$$

$$= \left(\frac{3}{4\pi \cdot k}\right)^{2/3} \cdot \left[\Gamma\left(\frac{5}{3}\right) - \Gamma^2\left(\frac{4}{3}\right)\right] \approx 0{,}0405 \cdot k^{-2/3}.$$

Среднеквадратическое отклонение:

$$\sigma \approx 0{,}201 \cdot k^{-1/3}.$$

4.9 Неравенства для вероятностей случайных величин и их характеристик

Неравенства, связывающие характеристики случайных величин, используются в тех случаях, когда расчет тех или иных характеристик сложен.

4.9.1 Неравенство Маркова

Пусть X — неотрицательная случайная величина и $M[X]$ — ее существующее математическое ожидание. Тогда для $\forall t > 0$

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

Докажем это неравенство, воспользовавшись определением математического ожидания, которое, по условию, конечно. Действительно,

$$M[X] = \int_{-\infty}^\infty xf(x)dx = \int_0^\infty xf(x)dx =$$

$$= \int_0^t xf(x)dx + \int_t^\infty xf(x)dx \geqslant \int_t^\infty xf(x)dx \geqslant$$

$$\geqslant t \cdot \int_t^\infty f(x)dx = t \cdot P(X > t).$$

4.9.2 Неравенство Чебышёва

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

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

Действительно, используя неравенство Маркова, получаем

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

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

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

4.9.3 Неравенство Хефдинга

Пусть $X_1, X_2, \ldots, X_n$ — независимые случайные величины, такие, что $M[X_i] = 0$ и $a_i \leqslant X_i \leqslant b_i$. Тогда для $\forall\varepsilon{>}0$ и $\forall t > 0$, [14]:

$$P\left(\sum_{i=1}^{n} X_i \geqslant \varepsilon\right) \leqslant \exp\Big\{-t \cdot \varepsilon\Big\} \cdot \prod_{i=1}^{n} \exp\Big\{t^2 \cdot \frac{(b_i-a_i)^2}{8}\Big\}.$$

Если $X_1, X_2, \ldots, X_n$ — независимые случайные величины, имеющие *распределение Бернулли* (*биномиальное распределение*, см. раздел 5.2) с параметром p, то выполняется неравенство Хефдинга:

$$P\Big(\big|\bar{x}-p\big| > \varepsilon\Big) \leqslant 2 \cdot \exp\Big\{-2n\varepsilon^2\Big\},$$

где $\bar{x}$ — среднее выборочное значение:

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

4.9.4 Неравенство Милла

Для случайной величины, имеющей *нормальный закон распределения* (см. раздел 5.14) с нулевым средним и единичной дисперсией ($X \sim N(0,1)$), выполняется неравенство Милла:

$$P\big(|X| > t\big) \leqslant \sqrt{\frac{2}{\pi}} \cdot \frac{\exp\left\{-\dfrac{t^2}{2}\right\}}{t}.$$

4.9.5 Неравенство Коши–Шварца

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

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

5 Основные законы распределения случайной величины

5.1 Распределение точечной массы

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

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

Плотность этого распределения:

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

Важно обратить внимание, что это дискретное распределение.

5.2 Биномиальное распределение

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

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

где $q = 1 - p$; $m = 0, \ldots, n$ — число появлений события A.

Биномиальное распределение однозначно задается двумя параметрами: количеством элементов выборки n и вероятностью p появления события A.

Математическое ожидание:

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

Дисперсия:

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

5.2.1 Понятие и использование производящей функции для вычисления характеристик распределений

Введем понятие *производящей функции*, которая является *преобразованием Лапласа* функции $f(x)$:

$$G_X(t) = M\Big[\exp\big\{tX\big\}\Big] = \int_{-\infty}^{+\infty} \exp\big\{tx\big\} \cdot f(x)dx.$$

Пусть проводится n опытов. Событие A появляется в каждом опыте с вероятностью p_i и не появляется с противоположной вероятностью $1-p_i$.

Пусть X — случайная величина, которая есть число появлений события A. Представим X как сумму величин X_i, каждое из которых принимает или значение 0 (событие A не появилось в i-м опыте), или 1 (событие A появилось в i-м опыте):

$$X = X_1 + X_2 + \cdots + X_n.$$

По определению математического ожидания дискретной случайной величины, производящая функция для события X_i есть

$$G_{X_i}(t) = M\Big[\exp\big\{tX_i\big\}\Big] = p_i \cdot \exp\big\{t \cdot 1\big\} + (1-p_i) \cdot \exp\big\{t \cdot 0\big\}.$$

Далее, для двух случайных величин X и Y по определению производящей функции следует, что производящая функция их суммы равна произведению их производящих функций:

$$G_{X+Y}(t) = G_X(t) \cdot G_Y(t).$$

Следовательно, для рассматриваемого случая

$$G_X(t) = \prod_{i=1}^{n} \Big(p_i \exp\{t\} + (1 - p_i) \Big).$$

Вычислим первую и вторую производные производящей функции при $t = 0$.

$$\frac{d}{dt} G_X(t)_{t=0} = M\left[\frac{d}{dt} \exp\{tX\} \right]_{\big| t=0} = M[X];$$

$$\frac{d^2}{dt^2} G_X(t)_{t=0} = M\left[\frac{d}{dt} (X \cdot \exp\{tX\}) \right]_{\big| t=0} = M[X^2].$$

Аналогично для высших моментов.

5.2.2 Вывод величины математического ожидания и дисперсии биномиального распределения с помощью производящей функции

Биномиальному распределению соответствует случай, когда $p_i = p$. Тогда математическое ожидание биномиального распределения есть

$$\frac{d}{dt} G_X(t)_{t=0} = n \left(p \cdot \exp\Big\{t\Big\} + (1-p) \right)^{n-1} \cdot p \cdot \exp\Big\{t\Big\}_{\big| t=0} = np.$$

Дисперсия может быть вычислена по формуле

$$D[X] =$$

$$= M[X^2] - \big(M[X]\big)^2 = \frac{d^2}{dt^2} G_X(t)_{\big| t=0} - \Big(\frac{d}{dt} G_X(t) \Big)^2_{\big| t=0} =$$

$$= \frac{d}{dt} \left(n \Big(p \cdot \exp\{t\} + (1-p) \Big)^{n-1} \cdot p \cdot \exp\{t\} \right)_{\big| t=0} -$$

$$- \big(np\big)^2 = np(1-p).$$

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, оставляя конечным и постоянным произведение $n \cdot p$, 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[10] δx равна $\lambda\delta x$, где λ — постоянная величина, то вероятность того, что в ограниченном интервале Δx событие A произойдет ровно k раз, дается распределением Пуассона:

$$p(x_k) = \frac{(\lambda \cdot \Delta x)^k}{k!} \exp\Big\{-\lambda \cdot \Delta x\Big\}.$$

Действительно,

$$p(x_k) = \frac{n(n-1)\dots(n-(k-1))}{k!}\left(\frac{\lambda \cdot \Delta x}{n}\right)^k\left(1-\frac{\lambda \cdot \Delta x}{n}\right)^{n-k} =$$

$$= \frac{(\lambda \cdot \Delta x)^k}{k!}\left(1\cdot\big(1-\frac{1}{n}\big)\cdot\big(1-\frac{2}{n}\big)\dots\big(1-\frac{k-1}{n}\big)\cdot\big(1-\frac{\lambda \cdot \Delta x}{n}\big)^{n-k}\right),$$

что в пределе больших n дает искомое выражение.

Математическое ожидание:

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

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

Дисперсия:

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

Для распределения Пуассона среднее число появлений события в разных выборках остается постоянным.

Распределение Пуассона можно использовать в качестве приближения для биномиального распределения (при малых $p_k \leqslant 0{,}10$),

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

ПРИМЕР. Число распадов радиоактивного вещества за время t.

ПРИМЕР. Число космических частиц, попадающих на поверхность площади S за время t.

ПРИМЕР. Понятие пуассоновского поля. Случайное поле точек называется *пуассоновским полем*, если выполняются следующие условия:

1. точки распределяются в поле статистически равномерно со средней плотностью λ (величина на единицу площади или на единицу объема);

2. точки попадают в непересекающиеся области независимо друг от друга;

3. 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:

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

где $a = S_g \cdot \lambda$ (для распределения на плоскости); $a = V_g \cdot \lambda$ (для распределения по объему). Примером такого поля

может служить система ICRF (международная небесная система отсчета, сформированная по далеким источникам, преимущественно квазарам).

5.4 Геометрическое распределение

Геометрическим распределением называется закон распределения числа независимых опытов (случайная величина X) с двумя исходами $\{A, \bar{A}\}$ (событие A или появляется, или не появляется) в одинаковых условиях

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

до первого появления события A:

$$p(x_k) = (1-p)^{k-1} \cdot p.$$

Другими словами, видоизменяется условие биномиального распределения: испытания заканчиваются, как только появляется событие A. Вероятность представляет собой геометрическую прогрессию с первым членом p и знаменателем $1-p$.

Математическое ожидание:

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

Дисперсия:

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

5.5 Гипергеометрическое распределение

Понятие *гипергеометрического распределения* сформулируем на примере. Пусть есть N изделий. Среди них M стандартных ($M < N$). Случайно выбирают n изделий

(отобранные изделия обратно не возвращают, т.е. формула Бернулли не применима). Случайная величина X — это количество испытаний, при котором есть ровно m стандартных деталей. Тогда распределение этой случайной величины есть четырехпараметрическое (N, M, m, n) дискретное гипергеометрическое распределение:

$$p(x_m) = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n}.$$

5.6 Показательное распределение

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

$$f(x) = \begin{cases} \lambda \cdot \exp\big\{-\lambda \cdot x\big\}, & \text{если } x \geqslant 0 \\ 0, & \text{если } x < 0. \end{cases}$$

Математическое ожидание:

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

Дисперсия:

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

Мода $x_l = 0$, а медиана не совпадает ни с модой, ни с математическим ожиданием и равна $x_m = \ln 2/\lambda$.

5.7 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:

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

Дисперсия:

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

Медиана x_m равна математическому ожиданию.

5.8 Распределение Вейбулла

Плотность *распределения Вейбулла* имеет вид:

$$f_{\alpha;\beta}(x) = \frac{\beta}{\alpha^\beta} x^{\beta-1} \exp\left\{-\left(\frac{x}{\alpha}\right)^\beta\right\}.$$

Ограничения на переменную и параметры следующие:

$$x \geqslant 0; \alpha, \beta > 0$$

При $\beta = 1$ распределение Вейбулла переходит в *экспоненциальное распределение*. При $\beta = 2$ — в *распределение Рэлея*:

$$f_\alpha(x) = \frac{2x}{\alpha^2} \exp\left\{-\frac{x^2}{\alpha^2}\right\}.$$

5.9 Гамма-распределение

Гамма-распределение — это распределение суммы $\alpha+1$ независимых случайных величин, каждая из которых имеет экспоненциальное распределение. Плотность распределения имеет вид

$$f_{\lambda;k}(x) = \frac{\lambda^k}{\Gamma(k)} x^{k-1} \exp\{-\lambda x\}; \;\; x > 0.$$

Здесь $\Gamma(k)$ — упомянутая в примере раздела 4.7.2. гамма-функция:

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

Для натуральных k

$$\Gamma(k+1) = k!; \;\; \Gamma(k+1) = k\Gamma(k).$$

При $\lambda = 1/2$ и $k = m/2$ гамма-распределение переходит в *распределение хи-квадрат* с m степенями свободы ($\chi^2(m)$). При $k = 1$ — в *показательное распределение.*

5.10 Бета-распределение

Плотность *бета-распределения* есть

$$f_{\alpha;\beta}(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}; \;\; \alpha > 0; \beta > 0.$$

Здесь $B(\alpha,\beta)$ есть бета-функция

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Через бета-распределение могут быть выражены многие как непрерывные, так и дискретные распределения.

5.11 Распределение Стьюдента

Распределение Стьюдента (*t-распределение*) определяется плотностью вероятности

$$f_k(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{\pi k}} \cdot \left(1 + \frac{x^2}{k}\right)^{-(k+1)/2},$$

где k — число степеней свободы.

Важно отметить, что случайная величина, имеющая распределение Стьюдента (обозначается T), есть отношение величины, имеющей стандартное нормальное распределение $X \sim N(0,1)$, к величине, связанной с распределением хи-квадрат:

$$T = \frac{X}{\sqrt{Y/k}},$$

где $Y \sim \chi^2(k)$ (k — число степеней свободы).

Распределение Стьюдента используется, в частности, для проверки равенства математических ожиданий двух выборок. При большом объеме выборки ($n > 30$) распределение Стьюдента переходит в нормальное распределение. Для сравнения выборок малых объемов используются непараметрические критерии (см. раздел 15).

5.12 Распределение Фишера

Распределение Фишера (распределение Фишера–Снедекора или *F-распределение* определяется плотностью вероятности

$$f_{f_1;f_2}(x) = \frac{\Gamma\left(\frac{f_1+f_2}{2}\right)}{\Gamma\left(\frac{f_1}{2}\right)\Gamma\left(\frac{f_2}{2}\right)} \cdot \left(\frac{f_1}{f_2}\right)^{f_1/2} \cdot \frac{x^{f_1/2-1}}{\left(1+\frac{f_1}{f_2}x\right)^{(f_1+f_2)/2}}.$$

Случайная величина, имеющая распределение Фишера (обозначается F), может быть представлена как

$$F = \frac{Y_1 \cdot f_2}{Y_2 \cdot f_1},$$

где $Y_1 \sim \chi^2(f_1)$; $Y_2 \sim \chi^2(f_2)$.

Распределение Фишера определено только для неотрицательных x и используется, в частности, для проверки равенства дисперсий двух выборок. Для сравнения выборок малых объемов используются непараметрические критерии (см. раздел 15).

5.13 Распределение Максвелла

Плотность *распределения Максвелла* имеет вид

$$f_\beta(x) = \begin{cases} \sqrt{\frac{2}{\pi}} \cdot \beta^{3/2} \cdot x^2 \cdot \exp\left\{ -\frac{\beta x^2}{2} \right\}, & \text{если } x > 0 \\ 0, & \text{если } x \leqslant 0. \end{cases}$$

В статистической физике параметр $\beta > 0$ определяется температурой и массой молекул. Распределением Максвелла, в частности, обладает случайная абсолютная скорость молекул идеального газа.

5.14 Нормальное распределение

Нормальное распределение (или *распределение Гаусса*) играет фундаментальную роль в теории ошибок в силу следующих причин:

1. нормальное распределение описывает распределение ошибок, возникающих при множестве малых независимых вкладов, носящих случайный произвольный характер;

2. некоторые функции случайной величины — например среднее значение — распределены асимптотически нормально даже тогда, когда исходная случайная величина не обладает нормальным распределением.

5.14.1 Основные понятия

Плотность вероятности для случайной величины X, имеющей нормальное распределение:

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

При больших по модулю значениях x плотность вероятности быстро симметрично спадает к нулю. Так, для $x = \pm 5$ плотность вероятности $f(x) \sim 10^{-6}$ (для $\mu = 0; \sigma^2 = 1$).

Функция распределения есть, соответственно,

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

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

$$M[X] = \mu;$$

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

В справедливости двух последних соотношений можно убедиться прямым вычислением. Действительно, по определению математического ожидания,

$$M[X] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left\{ \frac{-(t-\mu)^2}{2\sigma^2} \right\} t dt = \mu.$$

При вычислениях был использован табличный интеграл $\mathrm{erf}(x)$ (*функция ошибок*), с его помощью было получено

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left\{ \frac{-(t-\mu)^2}{2\sigma^2} \right\} dt = 1,$$

а также учитывалась нечетность подынтегральной функции и, следовательно, равенство нулю интеграла

$$\int_{-\infty}^{+\infty} \exp\left\{\frac{-(t-\mu)^2}{2\sigma^2}\right\}(t-\mu)dt = 0.$$

Дисперсия:

$$D[X] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left\{\frac{-(t-\mu)^2}{2\sigma^2}\right\}(t-\mu)^2 dt = \sigma^2.$$

Крутизна (*эксцесс*) нормально распределенной случайной величины X равна 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.

Скошенность (*асимметрия*) нормально распределенной случайно величины X также равна нулю:

$$\gamma_1 = \frac{\mu_3[X]}{\sigma^3[X]} = 0,$$

где $\mu_3[X]$ — центральный момент третьего порядка.

В двух последних равенствах можно убедиться прямым вычислением.

Ненулевые значения эксцесса и асимметрии используются как характеристики отклонения исследуемого распределения от нормального распределения.

Так, при $\gamma_1 > 0$ имеет место *правосторонняя асимметрия*, при $\gamma_1 < 0$ — *левосторонняя асимметрия*. Далее, при одном и том же параметре σ максимум плотности распределения с $\gamma_2 < 0$ ниже максимума плотности нормального распределения, которое, в свою очередь, ниже

максимума плотности распределения для распределения с $\gamma_2 > 0$.

Случайную величину X, распределенную по нормальному закону, обозначают

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

Для удобства работы с нормально распределенными величинами и для подсчета необходимых вероятностей с помощью статистических таблиц распределение приводят к стандартному виду следующим образом. Вводят замену переменной

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

Случайная величина U имеет *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

$$u = \frac{x - \mu}{\sigma}; \quad x = \sigma u + \mu;$$

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

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

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left\{ -\frac{t^2}{2} \right\} dt.$$

Эта функция задана таблично. Еще более удобно использовать *функцию Лапласа–Гаусса*, также заданную таблично, которая есть

$$\Phi_0(u) = \frac{1}{\sqrt{2\pi}} \int_0^u \exp\left\{-\frac{t^2}{2}\right\} dt.$$

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

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

5.14.2 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 указывает свойства и характеристики распределения выборочного среднего для выборки, имеющий произвольный закон распределение.

Эта теорема формулируется следующим образом. Пусть произвольная 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 равноточной 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

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

Здесь необходимо дать следующее пояснение. В задачах обработки данных считается, что величина x_i (i-я реализация случайной величины X) и величина $\bar{x}_i$ (i-я 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 поэтому 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 важнейшую формулу для оценки среднеквадратического отклонения выборочного среднего:

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

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

5.14.3 Доказательство центральной предельной теоремы

Для доказательства центральной предельной теоремы используем введенное в разделе 5.2.1 понятие *производящей функции*. Так, производящая функция нормально распределенной случайной величины $X \sim N(\mu, \sigma^2)$ есть

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

Первая производная производящей функции в нуле

$$\psi'(0) = \Big[\frac{d}{dt} M\Big[\exp\big\{tX\big\}\Big]\Big]_{\big|t=0} = M\Big[\frac{d}{dt}\exp\big\{tX\big\}\Big]_{\big|t=0} =$$

$$= M\Big[X\exp\big\{tX\big\}\Big]_{\big|t=0} = M[X].$$

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 $\psi_X(t)$ — производящая функция случайной величины X, то производящая функция случайной величины Y:

$$\psi_Y(t) = \exp\left\{bt\right\} \cdot \psi_X(at).$$

Если $X_1, X_2, \ldots, X_n$ — независимые случайные величины и $Y = \sum\limits_i X_i$, то

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

где $\psi_i(t)$ — 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 $\sum\limits_i Y_i$ 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

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

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

Воспользуемся известным пределом

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

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

Тогда

$$\xi(t) = \lim_{n\to\infty} \xi_n(t) = \exp\left\{\frac{t^2}{2}\right\}.$$

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

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

что и требовалось доказать.

5.14.4 Правило 3σ (трех сигм)

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:

$$P(\mu - l \leqslant X < \mu + l) = P\bigg(-\frac{l}{\sigma} \leqslant \frac{X - \mu}{\sigma} < \frac{l}{\sigma}\bigg) =$$

$$= \Phi_0\bigg(\frac{l}{\sigma}\bigg) - \Phi_0\bigg(-\frac{l}{\sigma}\bigg) = 2\Phi_0\bigg(\frac{l}{\sigma}\bigg).$$

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

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

Число l, вообще говоря, любое положительное число. 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 среднеквадратического (стандартного) отклонения — это правило 1σ (одной сигма).

Аналогично определяется правило 2σ (двух сигм):

$$P\Big(|X-\mu| < 2\sigma\Big) = 2\Phi_0(2) = 0{,}954$$

и правило 3σ (трех сигм):

$$P\Big(|X-\mu| < 3\sigma\Big) = 2\Phi_0(3) = 0{,}997.$$

Согласно последнему равенству, все значения случайной величины X, распределенной по нормальному закону, с вероятностью 99,7% укладываются в интервале $[\mu-3\sigma;\mu+3\sigma)$[11].

5.14.5 Таблица стандартного нормального распределения. Правила работы с таблицей

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

[11] 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.

(функции Лапласа–Гаусса) $\Phi_0(u)$:

$$\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\Big\{-\frac{t^2}{2}\Big\} dt; \tag{7}$$

$$\Phi_0(u) = \frac{1}{\sqrt{2\pi}} \int_{0}^{u} \exp\Big\{-\frac{t^2}{2}\Big\} dt. \tag{8}$$

Таблица функции $\Phi(u)$ содержит вычисленный интеграл (7) для каждого u. Таблица содержит вероятности того, что случайная величина

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

(где X — нормально распределенная случайная величина со средним μ и среднеквадратическим отклонением σ) принимает значения, меньше u, см. табл. 13.

Соответствующие величины u заданы с интервалом 0,1 в крайнем левом столбце табл. 13. Если требуется найти $\Phi(u)$ для u, заданных с лучшей точностью, до 0,01, используется первая строка табл. 13, 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 табл.13 величина u меняется от 0 до примерно 3,4, 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 $\Phi(3{,}4) = 0{,}9998$ и с ростом u только растет в силу своей монотонности.

Таблица 13

Функция распределения $\Phi(u)$ нормального закона $N(0,1)$.

u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Таблица для функции Лапласа–Гаусса $\Phi_0(u)$ содержит вычисленный интеграл (8) для каждого u и устроена полностью аналогично табл. 13 для функции $\Phi(u)$. Все вероятности могут быть получены из табл. 13 функции $\Phi(u)$, поскольку имеет место следующее соотношение между $\Phi(u)$ и $\Phi_0(u)$, указанное в разделе 5.14.1:

$$\Phi_0(u) = \Phi(u) - \frac{1}{2}.$$

Приведем примеры задач на нормальное распределение.

ПРИМЕР. Изготовлена цилиндрическая деталь диаметром D [9]. 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: $D < 39{,}85$ мм или $D > 40{,}05$ мм. Следует 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; $\beta = 40{,}05$ 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 табл. 13 функции $\Phi(u)$:

$$P(\bar{A}) = P\Big(D \in [39{,}85; 40{,}05]\Big) =$$

$$= \Phi\left(\frac{40{,}05 - 40{,}00}{0{,}05}\right) - \Phi\left(\frac{39{,}85 - 40{,}00}{0{,}05}\right) = \Phi(1) - \Phi(-3) =$$

$$\Phi(1) - 1 + \Phi(3) = 0{,}8413 - 1 + 0{,}9987 = 0{,}8400.$$

Вычислим средний процент забракованных деталей:

$$P(A) = 1 - 0{,}84 = 0{,}16 = 16\%.$$

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

Используем правило трех сигм для нахождения величины среднеквадратического отклонения случайной величины Δh:

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

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

$$P(|\Delta h - m_{\Delta h}| < 10) = 2\Phi_0(1) = 0{,}683.$$

5.15 Распределения, близкие к нормальному распределению

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

Например, кривая распределения параллаксов звезд ограничена справа и слева, в отличие от нормального распределения, поскольку все параллаксы больше нуля и не

существует очень больших параллаксов. Кроме того, с уменьшением параллакса число звезд растет.

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 являющихся, что позволяет аппроксимировать наблюдательные и экспериментальные данные более точными кривыми.

Рассмотрим плотность распределения

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

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 присутствуют введенные ранее в разделе 4.6 асимметрия:

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

и эксцесс:

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

Величина $\mu_k = M[(x-\mu)^k]$ есть k-й 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 $\sigma = \sigma(u)$.

Напомним, что если эмпирическое распределение содержит два максимума или более, то удобно представить его суммой двух распределений или более. Также заметим, что максимумы следует отличать от случайных выбросов, которые могут быть вызваны как ошибками при сборе данных, так и инструментальными ошибками.

5.16 Распределения, связанные с нормальным распределением

Среди наиболее часто используемых распределений, связанных с нормальным, отметим *распределение* χ^2 (или χ^2*-распределение*) и *log-нормальное распределение.*

5.16.1 Распределение χ^2 (хи-квадрат)

Распределение χ^2 с n степенями свободы (обозначается $\chi^2(n)$) есть распределение суммы квадратов n неза-

висимых случайных величин, имеющих стандартное нормальное распределение

$$Y = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n); \;\; X_i \sim N(0,1).$$

Математическое ожидание: $M[Y] = n$.
Дисперсия: $D[Y] = 2n$.
Асимметрия: $\gamma_1 = \sqrt{8}/n$.
Эксцесс: $\gamma_2 = 12/n$.

По центральной предельной теореме, при $n \to \infty$ случайная величина $Y \sim N(n, 2n)$.

Как было отмечено ранее в разделе 5.9, распределение $\chi^2(n)$ есть частный случай гамма-распределения, следовательно, функция плотности распределения хи-квадрат:

$$f(x) = f(x; \lambda = 1/2, k = n/2) = \frac{\lambda^k}{\Gamma(k)} x^{k-1} \exp\{-\lambda x\} =$$

$$= \frac{(1/2)^{n/2}}{\Gamma(1/2)} \cdot x^{n/2-1} \cdot \exp\{-\frac{1}{2}x\}.$$

Наконец, $\chi^2(2)$ — это экспоненциальное распределение с функцией плотности

$$f(x) = \frac{1}{2} \exp\{-\frac{1}{2}x\}.$$

5.16.2 Log-нормальное распределение

Пусть случайная величина $Y \sim N(0,1)$. Тогда случайная величина X такая, что $Y = \log X$, имеет *log-нормальное распределение* $X \sim \log N(0,1)$. Соответствующая функция плотности log-нормального распределения:

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

Это выражение легко получить по правилу вычисления распределения вероятности для функции от случайной величины, заметив, что

$$f(y(x)) = \frac{1}{\sqrt{2\pi}} \exp\Big\{-\frac{1}{2}\log^2 x\Big\},$$

$$\frac{1}{x} = \Big|\frac{dy(x)}{dx}\Big|.$$

Для log-нормального распределения математическое ожидание и дисперсия есть, соответственно,

$$M[X] = \exp\Big\{\mu + \frac{\sigma^2}{2}\Big\},$$

$$D[X] = \Big(\exp\big\{\sigma^2\big\} - 1\Big) \cdot \exp\big\{2\mu + \sigma^2\big\}.$$

Для моделирования log-нормального распределения сначала моделируется нормальное распределение, а затем его экспонента:

$$x_i \to \exp\big\{x_i\big\}.$$

Для моделирования самого нормального распределения можно действовать двумя способами:

1. воспользоваться центральной предельной теоремой, составив выборку из средних значений некоторых промежуточных выборок с произвольными законами распределения;

2. применить к равномерному распределению $\{y_j\}$, например, следующее преобразование

$$x_i = \sqrt{-2\log y_{j-1}} \cdot \cos 2\pi y_j \quad (j = 2, 3, \ldots, n).$$

6 Точечные и интервальные оценки

Цель математической статистики — указать методы, с помощью которых по данным выборки можно получить оценки параметров генеральной совокупности.

Пусть производится наблюдение какой-либо случайной величины или, что математические одно и то же, регистрируется последовательность наблюдательных данных. Если наблюдения ведутся на достаточно большом промежутке времени (в идеале — бесконечно долго), то по результатам наблюдений можно точно вычислить такие параметры как среднее, среднеквадратическое отклонение и среднеквадратическое отклонение среднего. Однако в реальности наблюдатель никогда не имеет дело с бесконечным набором наблюдений случайной величины (с генеральной совокупностью). Таким образом, параметры генеральной совокупности всегда остаются неизвестными. В распоряжении наблюдателя имеется только ограниченный набор данных (выборка), и только с помощью этого набора нужно можно получать 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 может являться среднее арифметическое элементов выборки.

Интервальная оценка указывает доверительный интервал для точечной оценки, т.е. насколько хороша эта оценка в рамках определенных критериев. Например, с ростом числа элементов выборки интервальная оценка должна становиться *уже*, поскольку чем больше выборка, тем больше информации и тем ближе оценка к истинному значению параметра. Интервальная оценка записывается как

$$J_\vartheta = \{\vartheta^* - \varepsilon; \vartheta^* + \varepsilon\},$$

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

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

$$\gamma = P(|\vartheta^* - \vartheta| < \varepsilon)$$

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:

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

где X — число опытов, в которых событие A произошло; n — число всех проведенных опытов (количество элементов выборки).

Пусть происходит повторение серий из n опытов каждая. Тогда X — случайная величина. Запишем X в виде

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

где

$$X_i(A) = \begin{cases} 1, & A \text{ появилось (вероятность } p) \\ 0, & A \text{ не появилось (вероятность } q = 1-p). \end{cases}$$

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,

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

Теперь построим интервальную оценку для p:

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

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:

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

Напомним, что ε — *точность оценки*; γ — *доверительная вероятность* (или *надежность*, или *достоверность*) оценки; $\alpha = 1-\gamma$ называется *уровнем значимости*, или *процентной точкой* (обозначается в процентах; например, для $\gamma = 0{,}95$ $\alpha = 5\%$); p^* — точечная оценка параметра p; $q^* = 1-p^*$.

Обозначим

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

Тогда

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

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

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

Квантиль распределения случайной величины (нормального, t-распределения Стьюдента, F-распределения Фишера, χ^2-распределения и др.) характеризует *критическое значение* для определения допустимых интервалов изменения оцениваемого параметра (математического ожидания, дисперсии, коэффициента корреляции и др.) этой случайной величины.

С помощью функции распределения стандартной нормальной величины $\Phi(u)$ квантиль u_γ вычисляется по формуле:

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

Получаем связи между точностью оценки, надежностью оценки и числом элементов выборки:

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

$$\varepsilon = \frac{u_\gamma\sqrt{p^* \cdot q^*}}{\sqrt{n}};$$

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

Окончательно, искомый доверительный интервал для p^* есть

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

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

ПРИМЕР. 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 (далее будем использовать только термин «доверительная вероятность») γ, можно вычислить точность этой точечной оценки ε, т.е. построить до-

верительный интервал точечной оценки. Для этого сначала вычислим квантиль u_γ с помощью функции Лапласа–Гаусса $\Phi_0(u)$:

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

Доверительная вероятность $\gamma = 0{,}95$

$$u_{0{,}95} = \Phi_0^{-1}(0{,}475) = \Phi^{-1}(0{,}975) = 1{,}96.$$

Здесь величины в скобках — вероятности, значения функции Лапласа–Гаусса и функции распределения стандартной нормальной величины соответственно. По значению последней по табл. 13 находится величина $u_{0{,}95}$.

Для удобства расчета квантилей нормально распределенной случайной величины пользуются табл. 14, сокращенным вариантом табл. 13: табл. 14 содержит только наиболее часто встречающиеся значения квантилей.

Найденное значение квантиля $u_\gamma = 1{,}96$ подставляем в выражение для точности оценки:

$$\varepsilon = \frac{u_\gamma\sqrt{p^*\cdot q^*}}{\sqrt{n}} = \frac{1{,}96\sqrt{0{,}70\cdot(1-0{,}70)}}{\sqrt{10}} = 0{,}28.$$

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

$$J_{p^*} = \{0{,}70-0{,}28;\, 0{,}70+0{,}28\} = \{0{,}42;\, 0{,}98\}.$$

ПРИМЕР. Ранее в разделе 4.9.3. рассматривалось неравенство Хефдинга: для независимых случайных величин, распределенных по закону Бернулли, $X_1, X_2, \ldots, X_n$ с параметром p

$$P\left(\left|\bar{x}-p\right| > \varepsilon\right) \leqslant 2\cdot\exp\left\{-2n\varepsilon^2\right\},$$

Таблица 14

Квантили u_γ нормального распределения $N(0,1)$.

$\Phi(u_\gamma) = \frac{1+\gamma}{2}$	0,9000	0,9500	0,9750	0,9900	0,9950	0,9990	0,9995
u_γ	1,282	1,645	1,960	2,326	2,576	3,090	3,291

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

$$P\left(\left|\bar{x}-p\right| > \varepsilon\right) \leqslant \alpha;\ \ \varepsilon = \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}.$$

6.1.1 Геометрическая интерпретация доверительного интервала оценки вероятности

Доверительный интервал, в котором лежит истинное значение оцениваемого параметра p определяется выражением

$$|p-p^*| < u_\gamma\sqrt{\frac{p\cdot(1-p)}{n}}.$$

Возведя обе части этого неравенства в квадрат, получим область, которая на плоскости в координатах (p, p^*) есть внутренняя часть эллипса [9]:

$$(p-p^*)^2 = \frac{u_\gamma^2}{n}\cdot p\cdot(1-p). \tag{9}$$

Прямая, параллельная оси Op и проходящая через фиксированную точку p^* (точечную оценку параметра p),

пересечет эллипс в двух точках. Длина этого сечения — доверительный интервал J_{p^*}. Точки пересечения в общем виде вычисляются из квадратного уравнения границы эллипса:

$$p_{1,2} = \frac{p^* + \dfrac{u_\gamma^2}{2n}}{1 + \dfrac{u_\gamma^2}{n}} \pm$$

$$\pm \frac{1}{1 + \dfrac{u_\gamma^2}{n}} \cdot \sqrt{\left(p^* + \frac{u_\gamma^2}{2n}\right)^2 - (p^*)^2 \cdot \left(1 + \frac{u_\gamma^2}{n}\right)}. \qquad (10)$$

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 выражение (10) приобретает простой вид:

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

что равносильно тому, как если бы в правой части уравнения (9) 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 μ^* математического ожидания μ принимают выборочное среднее (среднее арифметическое всех элементов выборки):

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

Возникает вопрос, можно ли в качестве точечной оценки математического ожидания выбрать другое среднее (геометрическое, гармоническое). Оказывается, среднее арифметическое является лучшей оценкой, и это можно доказать, например, с помощью *метода максимального правдоподобия (ММП)* или *метода наименьших квадратов (МНК)*.

6.2.2 Поиск точечной оценки математического ожидания методом максимального правдоподобия

На примере поиска точечной оценки математического ожидания случайной величины рассмотрим метод максимального правдоподобия (ММП).

Предположим, что вид закона распределения генеральной совокупности известен, но неизвестны параметры, численно 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. ММП заключается в том, что выбираются такие оценки параметров, которые дают максимальное значение плотности вероятности (другими словами, вероятность которых максимальна).

Если $x_i (i = 1, 2, \ldots, n)$, где все элементы не зависят друг от друга, составляют выборку с совместной плотностью вероятности $f(x_1, x_2, \ldots, x_n; \mu^*, \sigma^*)$, то эта плотность вероятность может быть представлена произведением плотностей всех x_i:

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

которая называется *функцией правдоподобия*.

Введем обозначение:

$$L = \ln\bigl[l(x_1, x_2, \ldots, x_n)\bigr].$$

(Максимумы функций l и L, очевидно, достигаются при одних и тех же значениях параметров.) Для того, чтобы μ^* и σ^* максимизировали функцию L, необходимо равенство нулю частных производных этой функции по этим параметрам:

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

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

Для нормального распределения

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

Далее, поскольку мы хотим оценить только математическое ожидание, только его обозначим «со звездочкой», μ^*. Все x_i — независимые, следовательно,

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

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

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

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

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

Таким образом, искомая оценка и есть среднее арифметическое элементов выборки:

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

что и требовалось доказать.

6.2.3 Поиск точечной оценки математического ожидания методом наименьших квадратов

Рассмотрим метод наименьших квадратов (МНК), также на примере 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 выражения

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

по μ^*. В точке минимума первая производная по μ с необходимостью равна 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, ММП и МНК дают одну и ту же точечную оценку математического ожидания — среднее арифметическое всех элементов выборки. Обратите внимание, что при использовании МНК 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[12]:

$$M[X_i] = \mu; \;\; D[X_i] = \sigma^2.$$

Тогда величина

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

есть точечная оценка математического ожидания с параметрами

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

[12] Здесь и далее, выборочные характеристики можно ассоциировать и с одной случайной величиной X_i, напрмер, $M[X_i]$; $D[X_i]$, формально подразумевая, что эти характеристики получены по серии каких-либо предыдущих наблюдений $\{X_{ij}\}$ $(j = 0, 1, 2, \ldots, m)$.

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

Таким образом, случайна величина

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

а доверительный интервал для μ^*

$$P\left(\left|\mu - \mu^*\right| < \varepsilon\right) = \gamma;$$

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

где точность оценки и квантиль определяются как

$$\varepsilon = \frac{u_\gamma \cdot \sigma}{\sqrt{n}};$$

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

Важно отметить, что в рассматриваемом случае среднеквадратическое отклонение σ известно, т.е. точность измерений должна быть априори 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 квантиля u_γ появится квантиль $t_{k,\gamma}$, имеющий *t-распределение Стьюдента с k* степенями свободы. Распределение Стьюдента имеет таблицы для своей функции распределения $T(t)$; при большом объеме выборки это распределение стремится к нормальному.

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

$$(\sigma^*)^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \mu^*)^2, \tag{11}$$

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

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

Формально поясним, откуда берется $(n-1)$ в выражении (11) для оценки дисперсии.

Для вывода этой формулы вычислим 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,

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

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

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.

Доверительный интервал для математического ожидания при неизвестной дисперсии имеет вид

$$P\left(\left|\mu^* - \mu\right| < \tilde{\varepsilon}\right) = \gamma;$$

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

где точность оценки есть

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

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

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

или, через процентный уровень значимости α,

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

где k — число степеней свободы распределения Стьюдента.

6.2.5 Использование распределения Стьюдента для построения интервальной оценки

Распределение Стьюдента и использование табл. 15 этого распределения разберем на примере [11].

ПРИМЕР. Задана выборка в виде вариационного ряда — измерения роста (в см) 10 человек:

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

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

Пусть случайная величина X — рост. Пусть эта случайная величина распределена по нормальному закону со средним $\mu^* = 167$ и неизвестной дисперсией $(\sigma^*)^2$:

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

Вычислим характеристики выборки.

Выборочное среднее есть среднее арифметическое всех элементов выборки: $\bar{x} = 172{,}4$. Это точечная оценка

Таблица 15

Квантили $t_{k,\boldsymbol{\gamma}} = T^{-1}\left(k, \dfrac{1+\boldsymbol{\gamma}}{2}\right)$ t-распределения Стьюдента.

Число степеней свободы k	$\frac{1+\gamma}{2}$ 0,750	0,900	0,950	0,975	0,990	0,995	0,999
1	1,000	3,078	6,314	12,796	31,821	63,657	318
2	0,816	1,886	2,920	4,303	6,965	9,925	22,3
3	0,765	1,638	2,353	3,182	4,541	5,841	10,2
4	0,741	1,533	2,132	2,776	3,747	4,604	7,173
5	0,727	1,476	2,015	2,571	3,365	4,032	5,893
6	0,718	1,440	1,943	2,447	3,143	3,707	5,208
7	0,711	1,415	1,895	2,365	2,998	3,499	4,785
8	0,706	1,397	1,860	2,306	2,896	3,355	4,501
9	0,703	1,373	1,833	2,262	2,821	3,250	4,297
10	0,700	1,372	1,812	2,228	2,764	3,169	4,144
11	0,697	1,363	1,796	2,201	2,718	3,106	4,025
12	0,695	1,356	1,782	2,179	2,681	3,055	3,930
13	0,694	1,350	1,771	2,160	2,650	3,012	3,852
14	0,692	1,345	1,761	2,145	2,624	2,977	3,787
15	0,691	1,341	1,753	2,131	2,602	2,947	3,733
16	0,690	1,337	1,746	2,120	2,583	2,921	3,686
17	0,689	1,333	1,740	2,110	2,567	2,898	3,646
18	0,688	1,330	1,734	2,101	2,552	2,878	3,610
19	0,688	1,328	1,729	2,093	2,539	2,861	3,579
20	0,687	1,325	1,725	2,086	2,528	2,845	3,552
21	0,686	1,323	1,721	2,080	2,518	2,831	3,527
22	0,686	1,321	1,717	2,074	2,508	2,819	3,505
23	0,685	1,319	1,714	2,069	2,500	2,807	3,485
24	0,685	1,318	1,711	2,064	2,492	2,797	3,467
25	0,684	1,316	1,708	2,060	2,485	2,787	3,450
30	0,683	1,310	1,697	2,042	2,457	2,750	3,385
40	0,681	1,303	1,684	2,021	2,423	2,704	3,307
60	0,679	1,296	1,671	2,000	2,390	2,660	3,232
120	0,677	1,289	1,658	1,980	2,358	2,617	3,160
∞	0,674	1,282	1,645	1,960	2,326	2,576	3,090

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 отличной от 0. Поскольку дисперсия априори не известна, надо сначала ее оценить при помощи выборочной дисперсии s^2:

$$(\sigma^*)^2 = s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \sum_{i=1}^{10} \frac{(x_i - 172{,}4)^2}{9} = 62{,}9;$$

$$s \approx 7.93.$$

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

$$\frac{s}{\sqrt{n}} = \sqrt{\frac{62{,}9}{10}} \approx 2{,}51.$$

По аналогии с z введем величину t:

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

Величина t служит критерием проверки, и нам необходимо определить закон ее распределения и вычислить 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 распределение $\chi^2(k)/k$ с $k = n-1$ степенями свободы.

Таким образом, t — функция двух случайных величин, чье распределение известно, а значит, само t тоже может быть вычислено по известным правилам. Величина t имеет распределение, называемое распределением Стьюдента с числом степеней свободы $k = n-1$ (таким же, как и у соответствующего $\chi^2(k)$-распределения). Из табл. 15 внимательный читатель заметит, что при большом количестве степеней свободы $k \to \infty$, распределение Стьюдента стремится к нормальному распределению:

$$t_{\infty,\gamma} = T^{-1}\Big(\infty, \frac{1+\gamma}{2}\Big) = u_\gamma = \Phi^{-1}\Big(\frac{1+\gamma}{2}\Big).$$

В нашем примере доверительная вероятность оценки есть

$$P\Big(|\mu^* - \bar{x}| < \tilde{\varepsilon}\Big) =$$

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

Подставляя найденные величины

$$\mu^* = 167{,}0; \;\; \bar{x} = 172{,}4; \;\; s/\sqrt{n} = 2{,}51; \;\; n = 10,$$

получаем

$$P\left(\frac{|167{,}0 - 172{,}4|}{2{,}51} < t_{9,\gamma}\right) =$$

$$= P\big(172{,}4 - 2{,}51 \cdot t_{9,\gamma} < 167{,}0 < 172{,}4 + 2{,}51 \cdot t_{9,\gamma}\big) = \gamma.$$

Осталось задать доверительную вероятность γ, вычислить $t_{9,\gamma}$ с помощью табл. 15 и проверить выполнение неравенства. Согласно стандартным рекомендациям, зададимся доверительной вероятностью $\gamma = 0{,}9$. Табличное

значение $t_{9;0,9} = T^{-1}(9;(1+0{,}9)/2) = T^{-1}(9;0{,}95) = 1{,}833$, т.е. с вероятностью 0,9 должно быть $|t_{9;0,9}| \leqslant 1{,}8331$.

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:

$$\frac{|167{,}0 - 172{,}4|}{2{,}51} = 2{,}15 > 1{,}833.$$

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, есть

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

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 (ММП) 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, [6]), обладающих разными точностями и полезные в прикладных задачах для быстрых расчетов. В качестве примера приведем простую линейную *оценку Даутона* для средеквадратического отклонения, которая при малых выборках $n \leqslant 10$ дает 94% эффективности по сравнению с оценкой ММП:

$$\sigma^* = \frac{1{,}77245}{n\cdot(n-1)}\sum_{i=1}^{n} x_i\cdot\big(2\cdot i - n - 1\big).$$

6.3.2 Интервальная оценка дисперсии

Пусть X_i — независимые нормально распределенные случайные величины с известным математическим ожиданием $M[X_i] = \mu$ и неизвестной истинной дисперсией $D[X_i] = \sigma^2$ [9].

В качестве точечной оценки дисперсии примем выражение

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

Для построения *интервальной оценки дисперсии* используется χ^2-распределение и соответствующая статистика

$$\chi^2(n) = \frac{n \cdot (\sigma^*)^2}{(\sigma)^2} = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2. \qquad (12)$$

Напомним, здесь и далее термином *«статистика»* обозначается величина, которая имеет заданный закон распределения. Для обозначения часто используется само название распределения: например, случайная величина X, имеющая распределение хи-квадрат с n степенями свободы, для удобства может быть обозначена просто как $\chi^2(n)$. Статистика — это неслучайное число, которое определяется реализациями случайной величины, в данном случае, x_i. Величина u (или u_γ), в данном примере,

$$u_i = \frac{x_i - \mu}{\sigma} \sim N\big(0, 1\big),$$

всегда обозначает статистику стандартного нормального распределения. Аналогично вводятся термины: t (или $t_{n,\gamma}$) — статистика для распределения Стьюдента; F (или $F_\gamma(k_1, k_2)$) — статистика для распределения Фишера.

Сначала поясним суть метода построения интервальной оценки.

Для интервальной оценки математического ожидания (см. раздел 6.2.4) использовалось соотношение вида

$$\gamma = P(|\mu - \mu^*| < \varepsilon)$$

и строилась величина

$$\frac{\mu - \mu^*}{\sigma/\sqrt{n}} \sim N(0, 1)$$

с целью использовать в качестве критерия известную табличную статистику u_γ, поскольку по центральной предельной теореме (см. раздел 5.14.2) среднее по выборке (μ) имеет нормальное распределение со средним μ^* и дисперсией σ^2/n.

Аналогично для дисперсии нас должна была бы интересовать вероятность

$$\gamma = P(|\sigma^2 - (\sigma^*)^2| < \varepsilon),$$

что однако не может быть вичислено с помощью закона нормального распределения, и потому статистика u_γ не может быть использована. Существует другая табличная статистика, которую можно использовать, но уже не для разности, а для отношения оценки и неизвестного истинного значения дисперсии:

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

где $(\sigma^*)^2$ — точечная оценка дисперсии; $\bar{x}$ — среднее арифметическое выборки; σ^2 — истинная неизвестная дисперсия.

Построенная величина имеет распределение $\chi^2(n)$ (12).

Таким образом, удалось связать истинную неизвестную дисперсию с известным табличным распределением:

$$\sigma^2 = n \cdot \frac{(\sigma^*)^2}{\chi^2(n)}.$$

Кроме того, интервал для дисперсии должен накрывать истинное значение дисперсии с доверительной вероятностью γ. Вероятность попадания случайной величины

$X \sim \chi^2(n)$ в интервал $[\alpha, \beta]$ есть[13]

$$P(X \in [\alpha, \beta]) = \int_{\alpha}^{\beta} f(x)dx,$$

где

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

Функция плотности $f(x)$ распределения хи-квадрат, очевидно, не является симметричной относительно оси ординат (в отличие от симметричной функции плотности стандартного нормального закона). Следовательно, доверительный интервал для оценки дисперсии J_{σ^*} нужно построить так, чтобы вероятность попадания случайной величины слева и справа от концов отрезка $[\alpha, \beta]$ была одинаковой и равной $(1-\gamma)/2$, где γ — доверительная вероятность оценки дисперсии:

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

Точки α и β

$$\alpha = \chi^2_{p_\alpha}(n);$$

$$\beta = \chi^2_{p_\beta}(n)$$

определяются по известным вероятностям p_α и p_β по табл. 16 распределения хи-квадрат:

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

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

[13] См. сноску 11.

Из определения p_α и p_β, очевидно, следует, что они имеют смысл не доверительных вероятностей, а процентных точек[14].

Доверительный интервал для дисперсии есть

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

где

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

есть точечная оценка дисперсии; μ априори известно. Этот доверительный интервал содержит (накрывает) неизвестную искомую дисперсию с заданной доверительной вероятностью γ. Точность оценки дисперсии ε:

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

ПРИМЕР. Производилась оценка дисперсии случайного параметра X по результатам 20 испытаний [9]. Результат статистической обработки значений $x_1, x_2, \ldots, x_{20}$ оказался равным $(\sigma^*)^2 = 16$, и среднее значение X было априори известно. Нужно определить интервальную оценку истинного значения дисперсии σ^2 при заданной доверительной вероятности $\gamma = 0{,}95$.

Для того, чтобы найти границы α и β для вычисления доверительного интервала, определим вероятности, которые по построению есть процентные точки p_α, p_β:

$$p_\alpha = \frac{1 + \gamma}{2} = 97{,}5\%;$$

[14] Напомним, что для доверительной вероятности γ процентная точка (в %) есть $\alpha = 1 - \gamma$.

$$p_\beta = \frac{1-\gamma}{2} = 2{,}5\%.$$

По табл. 16 χ^2-распределения найдем

$$\alpha = \chi^2_{p_\alpha=97{,}5\%}(20) = \chi^2_{1-0{,}975}(20) = \chi^2_{0{,}025}(20) = 9{,}591;$$

$$\beta = \chi^2_{p_\beta=2{,}5\%}(20) = \chi^2_{1-0{,}025}(20) = \chi^2_{0{,}975}(20) = 34{,}170.$$

Теперь вычислим границы доверительного интервала для оценки дисперсии:

$$J_{\sigma^*} = \left\{ \frac{20 \cdot 16}{34{,}170}; \frac{20 \cdot 16}{9{,}591} \right\} = \{9{,}36; 33{,}36\}.$$

Истинное (неизвестное) значение дисперсии σ^2 с вероятностью 95% накрывается отрезком $\{9{,}36; 33{,}36\}$. Доверительный интервал не является симметричным относительно точечной оценки $(\sigma^*)^2 = 16$, поэтому точность оценки дисперсии ε можно определить только приближенно по формуле

$$\varepsilon = \frac{1}{2} \cdot \left(\frac{n \cdot (\sigma^*)^2}{\alpha} - \frac{n \cdot (\sigma^*)^2}{\beta} \right) = \frac{33{,}36 - 9{,}36}{2} = 12{,}0.$$

Рассмотрим, как изменится интервальная оценка дисперсии, если математическое ожидание априори не известно.

Если математическое ожидание случайной величины X неизвестно, то его, как и дисперсию, нужно определять по выборке:

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

В этом случае за точечную оценку дисперсии принимается

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

Таблица 16

Некоторые значения статистики $\chi^2(k)$ для разных значений доверительной вероятности γ.

	Доверительная вероятность/процентная точка							
Число степеней свободы k	0,01 99%	0,025 97,5%	0,05 95%	0,1 90%	0,9 10%	0,95 5%	0,975 2,5%	0,99 1%
1	0,0315	0,0398	0,0239	0,0158	2,706	3,841	5,024	6,635
2	0,0201	0,0506	0,103	0,211	4,605	5,991	7,378	9,210
5	0,554	0,831	1,145	1,610	9,236	11,070	12,832	15,086
10	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209
15	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578
20	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566
50	29,707	32,357	34,764	37,689	63,167	67,505	71,420	76,154
100	70,065	74,222	77,929	82,358	118,498	124,342	129,561	135,807

Для того, чтобы получить интервальную оценку дисперсии, построим статистику $\chi^2(n-1)$:

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

где, как и в предыдущем случае, σ^2 — неизвестное истинное значение дисперсии. Число степеней свободы такой статистики есть $(n-1)$. Интервальная оценка дисперсии при неизвестном математическом ожидании строится аналогично рассмотренной выше, но концы доверительного интервала $\tilde{\alpha}$, $\tilde{\beta}$ определяются теперь как

$$\tilde{\alpha} = \chi^2_{p_\alpha}(n-1);\ \ \tilde{\beta} = \chi^2_{p_\beta}(n-1),$$

а концы доверительного интервала вычисляются по формуле

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

Точность оценки дисперсии при неизвестном математическом ожидании вычисляется как

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

Доверительный интервал для дисперсии, получаемый в условиях незнания математического ожидания, становится шире, чем доверительный интервал для дисперсии при известном математическом ожидании. Точность в первом случае хуже, чем во втором. Этот результат довольно очевиден, т.к. любая дополнительная информация способствует получению более точных оценок искомых неизвестных величин.

6.4 Сравнение дисперсий двух выборок нормальной генеральной совокупности

Пусть имеется две нормально распределенные случайные выборки:

$$X = \{x_1, x_2, \ldots, x_n\} \sim N(\mu_x, \sigma_x^2);$$

$$Y = \{y_1, y_2, \ldots, y_m\} \sim N(\mu_y, \sigma_y^2).$$

Для проверки того, можно ли с некоторой доверительной вероятностью γ считать равными дисперсии двух выборок, $\sigma_x^2 = \sigma_y^2$, используется соответствующий квантиль, или критическое значение, называемое F-статистикой, $F_\gamma(k_1, k_2)$ (распределение Фишера). По данным двух выборок вычисляется величина

$$\tilde{F} = \tilde{F}_{(\gamma+1)/2}(n-1, m-1) = \frac{s_1^2}{s_2^2} = \frac{\chi_\gamma^2(n-1)/(n-1)}{\chi_\gamma^2(m-1)/(m-1)}, \quad (13)$$

Таблица 17

Некоторые значения $F_{\gamma}(k_1, k_2)$ распределения Фишера. Верхняя цифра — 5%-е значения ($\gamma = 0{,}95$). Нижняя цифра — 1%-е значения ($\gamma = 0{,}99$).

	k_1								
k_2	1	2	4	5	7	8	9	10	11
1	161	200	225	230	237	239	241	242	243
	4052	4999	5625	5764	5928	5981	6022	6056	6082
2	18,51	19,00	19,25	19,30	19,36	19,37	19,38	19,39	19,40
	98,49	99,01	99,25	99,30	99,34	99,36	99,38	99,40	99,41
4	7,71	6,94	6,39	6,26	6,09	6,04	6,00	5,96	5,93
	21,20	18,00	15,98	15,52	14,98	14,80	14,66	14,54	14,45
5	6,61	5,79	5,19	5,05	4,88	4,82	4,78	4,74	4,70
	16,26	13,27	11,39	10,97	10,45	10,27	10,15	10,05	9,96
7	5,59	4,74	4,12	3,97	3,79	3,73	3,68	3,63	3,60
	12,25	9,55	7,85	7,46	7,00	6,84	6,71	6,62	6,54
8	5,32	4,46	3,84	3,69	3,50	3,44	3,39	3,31	3,31
	11,26	8,65	7,01	6,63	6,19	6,03	5,91	5,82	5,74
9	5,12	4,26	3,63	3,48	3,29	3,23	3,18	3,13	3,10
	10,56	8,02	6,42	6,06	5,62	5,47	5,35	5,26	5,18
10	4,96	4,10	3,48	3,33	3,14	3,07	3,02	2,97	2,94
	10,04	7,56	5,99	5,64	5,12	5,06	4,95	4,85	4,78
11	4,84	3,98	3,36	3,20	3,01	2,95	2,90	2,86	2,82
	9,85	7,20	5,67	5,32	4,88	4,74	4,63	4,54	4,46

которая далее сравнивается с табличной статистикой $F_{(\gamma+1)/2}(n-1, m-1)$, см. табл. 17–18.

Выражение (13) представляет собой связь распределения Фишера и распределения хи-квадрат.

Отношение оценок дисперсий всегда берется большая к меньшей, чтобы оно было больше единицы. Критический интервал определяется по табл. 17–18, где размеры выборок определяют соответствующее число степеней свободы табличных функций:

$$\Big\{F_{1,(\gamma+1)/2}(n-1, m-1); F_{2,(\gamma+1)/2}(n-1, m-1)\Big\},$$

Таблица 18

Некоторые значения $F_{\gamma}(k_1, k_2)$ распределения Фишера для больших значений степеней свободы. Верхняя цифра — 5%-е значения ($\gamma = 0{,}95$). Нижняя цифра — 1%-е значения ($\gamma = 0{,}99$)

	k_1			
k_2	50	75	100	1000
50	1,60	1,55	1,52	1,44
	1,4	1,86	1,82	1,68
70	1,53	1,47	1,45	1,35
	1,82	1,74	1,69	1,53
100	1,48	1,42	1,39	1,28
	1,73	1,64	1,59	1,43
1000	1,36	1,30	1,26	1,08
	1,54	1,44	1,38	1,11

причем

$$F_1 = F_{1,(\gamma+1)/2}(n-1, m-1) = \frac{1}{F_{(\gamma+1)/2}(m-1, n-1)};$$

$$F_2 = F_{2,(\gamma+1)/2}(n-1, m-1) = F_{(\gamma+1)/2}(n-1, m-1).$$

Другими словами, если

$$F_1 < \tilde{F} < F_2,$$

то дисперсии считаются одинаковыми, иначе — разными.

Рассмотрим пример на сравнение двух дисперсий и работу с таблицей распределения Фишера.

ПРИМЕР. В ходе проверок фиксировались отклонения показаний высотометров от точного значения высоты [9]. Результаты отклонений приведены в табл. 19.

Таблица 19

Точность работы двух высотомеров.

№ наблюдения i	Отклонение высотомера №1 (м)	Отклонение высотомера №2 (м)
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	

Требуется сравнить дисперсии двух выборок (сравнить точности двух высотомеров). Доверительная вероятность $\gamma = 0{,}9$ (процентная точка $\alpha = 1 - 0{,}9 = 10\%$).

Вычислим точечные оценки математических ожиданий и дисперсий величин

$$X = \{x_1, x_2, \ldots, x_{12}\};\ \ Y = \{y_1, y_2, \ldots, y_9\}.$$

Точечные оценки математического ожидания:

$$\bar{x} = \frac{1}{12}\sum_{i=1}^{12} x_i = 4{,}5;\ \ \bar{y} = \frac{1}{9}\sum_{i=1}^{9} y_i = 4{,}7.$$

Точечные оценки дисперсий:

$$s_x^2 = \frac{1}{12-1}\sum_{i=1}^{12}(x_i - \mu_x)^2 = 451{,}9;$$

$$s_y^2 = \frac{1}{9-1}\sum_{i=1}^{9}(y_i - \mu_y)^2 = 332{,}0.$$

Значение статистики Фишера есть отношение большей оценки дисперсии к меньшей:

$$\tilde{F} = \frac{451{,}9}{332{,}0} = 1{,}36.$$

Число степеней свободы табличной статистики есть 11 и 8 соответственно; $\gamma = 0{,}9$. Следовательно,

$$F_{2;(0,9+1)/2}(11;8) = F_{0,95}(11;8) = 3{,}31;$$

$$F_{1;(0,9+1)/2}(11;8) = \frac{1}{F_{0,95}(8;11)} = \frac{1}{2{,}95} = 0{,}34.$$

Окончательно получаем верное неравенство

$$0{,}34 < 1{,}36 < 3{,}31.$$

Следовательно, дисперсии двух выборок равны с доверительной вероятностью 0,9.

6.5 Сравнение математических ожиданий двух выборок нормальной генеральной совокупности

Пусть $\{x_i\}$ $(i = 1, 2, \ldots n)$; $\{y_i\}$ $(i = 1, 2, \ldots m)$ — две независимые выборки из нормальных генеральных совокупностей с априори известными, но не равными дисперсиями σ_x^2 и σ_y^2. Критерий оценки равенства математических ожиданий двух выборок μ_x и μ_y — нормированный на дисперсии оценок модуль разности

$$\frac{|\mu_x^* - \mu_y^*|}{\sqrt{\dfrac{\sigma_x^2}{n} + \dfrac{\sigma_y^2}{m}}} \sim N(0,1),$$

где по центральной предельной теореме

$$\sigma_x^2/n = D[\mu_x^*];\ \ \sigma_y^2/m = D[\mu_y^*].$$

Вычисленный критерий сравнивается с нормальной статистикой u_γ, и в случае, если он оказывается меньше, математические ожидания считаются равными. Если дисперсии априори не известны, то используется критерий Стьюдента, и задача усложняется. Однако на практике достаточно пользоваться простой вышеприведенной формулой.

6.6 Оценивание параметров угловых случайных величин

Кратко остановимся на важном для астрономических приложений вопросе оценивания математического ожидания и дисперсии угловых случайных величин [8]. Угловые случайные величины возникают при исследовании данных, распределенных на небесной сфере, а также при исследовании периодических процессов.

Пусть выборка n элементов — это Z_i, точки на окружности единичного радиуса ($i = 1, \ldots, n$). Тогда точечная оценка математического ожидания Z_i, ассоциированного с *математическим ожиданием угловой величины*, есть *выборочное круговое среднее направление* для углов φ_i, которое определяется как направление суммы векторов $\overline{OZ}_i$. Здесь O — центр единичной окружности, на которой лежат точки Z_i; углы φ_i задают положения этих точек в декартовой системе координат как $(\cos\varphi_i, \sin\varphi_i)$.

Другими словами, выборочное круговое среднее направление φ_m удовлетворяет равенствам

$$C_1 = \sqrt{C_1^2 + C_2^2} \cdot \cos\varphi_m;$$

$$C_2 = \sqrt{C_1^2 + C_2^2} \cdot \sin \varphi_m,$$

где C_1 и C_2 имеют смысл декартовых координат центра масс системы всех точек Z_i и определяются как

$$C_1 = \frac{1}{n} \sum_{i=1}^{n} \cos \varphi_i; \quad C_2 = \frac{1}{n} \sum_{i=1}^{n} \sin \varphi_i.$$

Если точки Z_i распределены по окружности равномерно, то длина вектора центра масс, которая называется *выборочной результирующей длиной* $\sqrt{C_1^2 + C_2^2}$, стремится к нулю. Если точки Z_i совпадают друг с другом, то длина вектора центра масс стремится к единице.

Выборочная круговая дисперсия направлений φ_i — это величина

$$V(\varphi_m) = \frac{1}{n} \sum_{i=1}^{n} \big(1 - \cos(\varphi_i - \varphi_m)\big) = 1 - \sqrt{C_1^2 + C_2^2},$$

не зависящая от выбора начала отсчета углов.

Выборочное круговое стандартное отклонение — это величина

$$s_\varphi = \sqrt{-2 \ln(1 - V(\varphi_m))} = \sqrt{-2 \ln \left(\sqrt{C_1^2 + C_2^2} \right)}.$$

Для малых значений выборочной круговой дисперсии верно соотношение, аналогичное для обычных дисперсии и стандартного отклонения случайной величины X ($D[X]$ и $s.d.$):

$$s_\varphi \approx \sqrt{2V(\varphi_m)}.$$

Все сказанное об оценках математического ожидания и дисперсии случайных углов может быть обобщено на двумерный случайный угловой вектор, когда положение точки задается уже двумя углами. Например, на небесной сфере — галактической долготой l и широтой b [8].

7 Перенос ошибок

7.1 Матрица ошибок

Часто результат эксперимента представляет собой функцию от нескольких 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 функции от x_r: $y = y(x_1, x_2, \ldots, x_m)$.

Для набора m случайных величин X_r $(r = 1, 2, \ldots, m)$ можно ввести *матрицу ошибок* $M_E(x)$, см. табл. 20.

Эта матрица 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 $(r = 1, 2, \ldots, m)$ есть линейные функции переменных $x_s (s = 1, 2, \ldots, n)$:

$$y_r = a_{r1}x_1 + a_{r2}x_2 + \cdots + a_{rn}x_n = \sum_{s=1}^{n} a_{rs}x_s$$

или в матричной форме

$$y = Ax.$$

Здесь a_{rs} — 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 $\bar{x}_s$.

Таблица 20

Элементы матрицы ошибок $M_E(x)$, записанные в виде таблицы.

$M[(x_1-\mu_1)^2]$	$M[(x_1-\mu_1)(x_2-\mu_2)]$	$\dots$	$M[(x_1-\mu_1)(x_r-\mu_r)]$
$M[(x_2-\mu_2)(x_1-\mu_1)]$	$M[(x_2-\mu_2)^2]$	$\dots$	$M[(x_2-\mu_2)(x_r-\mu_r)]$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$M[(x_r-\mu_r)(x_1-\mu_1)]$	$M[(x_r-\mu_r)(x_2-\mu_2)]$	$\dots$	$M[(x_r-\mu_r)^2]$

Если $M_E(x)$ — матрица ошибок величин x_s, тогда матрица ошибок для функций y_r:

$$M_E(y) = A \cdot M_E(x) \cdot A^T.$$

Докажем последнее равенство.

Пусть математическое ожидание всех X_r одинаковое: $\mu_1 = \dots = \mu_r = \mu$. Рассмотрим для простоты случай двух случайных величин x_1 и x_2, и две функции y_1 и y_2:

$$y_1 = a_{11}x_1 + a_{12}x_2; \;\; y_2 = a_{21}x_1 + a_{22}x_2.$$

Соответствующая матрица:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

Матрица ошибок:

$$M_E(x) =$$

$$= \begin{pmatrix} M[(x_1-\mu)^2] & M[(x_1-\mu)(x_2-\mu)] \\ M[(x_2-\mu)(x_1-\mu)] & M[(x_2-\mu)^2] \end{pmatrix}.$$

Тогда

$$B = A \cdot M_E(x) \cdot A^T =$$

$$= \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \times$$

$$\begin{pmatrix} M[(x_1-\mu)^2] & M[(x_1-\mu)(x_2-\mu)] \\ M[(x_2-\mu)(x_1-\mu)] & M[(x_2-\mu)^2] \end{pmatrix} \times$$

$$\begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix}.$$

Распишем первый элемент полученной квадратной матрицы B:

$$b_{11} = \big(a_{11}M[(x_1-\mu)^2] + a_{12}M[(x_1-\mu)(x_2-\mu)]\big) \cdot a_{11} +$$

$$+\big(a_{11}M[(x_1-\mu)(x_2-\mu)] + a_{12}M[(x_2-\mu)^2]\big) \cdot a_{12} =$$

$$= (a_{11})^2 M[(x_1-\mu)^2] + (a_{12})^2 M[(x_2-\mu)^2] +$$

$$+2a_{11}a_{12}M[(x_1-\mu)(x_2-\mu)].$$

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

$$\mu_{y1} = (a_{11} + a_{12})\mu;$$

$$\mu_{y2} = (a_{21} + a_{22})\mu,$$

распишем первый элемент:

$$M_E(y)_{11} = M[(a_{11}x_1 + a_{12}x_2 - (a_{11} + a_{12})\mu)^2] =$$

$$= M[(a_{11}(x_1-\mu) + a_{12}(x_2-\mu))^2] = (a_{11})^2 M[(x_1-\mu)^2] +$$

$$+(a_{12})^2 M[(x_2-\mu)^2] + 2a_{11}a_{12}M[(x_1-\mu)(x_2-\mu)] = b_{11}.$$

Аналогично поэлементно доказывается равенство матрицы B и $M_E(y)$, что и требовалось доказать.

Полученный результат можно легко обобщить на случай нелинейных функций $y_r = f_r(x_1, \ldots, x_s, \ldots, x_n)$, предположив, что функция f_r мало меняется в области, ограниченной среднеквадратическим отклонением от ее среднего значения.

Другими словами, с точностью до членов первого порядка разложение данной функции в ряд Тейлора имеет вид

$$y_r = f_r(x_1, \ldots, x_n) = f_r(\bar{x}_1, \ldots, \bar{x}_n) + \sum_{s=1}^{n} \left\{ (x_s - \bar{x}_s) \frac{\partial f_r}{\partial x_s} \Big|_{\bar{x}} \right\}.$$

Здесь $\bar{x}_s$ — среднее значение x_s; $\bar{x} = \{\bar{x}_1, \ldots, \bar{x}_n\}$. Тогда оценка величины y_r

$$y_r^* = f_r(\bar{x}_1, \ldots, \bar{x}_n),$$

а элементами матрицы ошибок являются

$$[M_E(y)]_{rs} = \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ (x_i - \bar{x}_i)(x_j - \bar{x}_j) \frac{\partial f_r}{\partial x_i} \Big|_{\bar{x}} \frac{\partial f_s}{\partial x_j} \Big|_{\bar{x}} \right\}.$$

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

$$[F]_{rs} = \frac{\partial f_r}{\partial x_s} \Big|_{\bar{x}}.$$

Далее рассмотрим применение матрицы ошибок для частных случаев функциональных зависимостей.

7.2 Отношение двух случайных величин

В качестве примера применения полученных формул рассмотрим одномерный случай *отношения двух случайных величин* x_1 и x_2.

Среднее значение этого отношения:

$$\bar{y} \approx \frac{\bar{x}_1}{\bar{x}_2}.$$

Матрица ошибок для $\bar{x}_1$ и $\bar{x}_2$:

$$M_E(x) = \begin{pmatrix} s_1^2 & s_1 s_2 q_{12} \\ s_1 s_2 q_{12} & s_2^2 \end{pmatrix},$$

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):

$$M_E(y) = \begin{pmatrix} 1/\bar{x}_2 & -\bar{x}_1/\bar{x}_2^2 \end{pmatrix} \begin{pmatrix} s_1^2 & s_1 s_2 q_{12} \\ s_1 s_2 q_{12} & s_2^2 \end{pmatrix} \begin{pmatrix} 1/\bar{x}_2 \\ -\bar{x}_1/\bar{x}_2^2 \end{pmatrix}.$$

После перемножения получаем ошибку величины $\bar{y}$:

$$s_{\bar{y}}^2 = \frac{\bar{x}_1^2}{\bar{x}_2^2} \left\{ \frac{s_1^2}{\bar{x}_1^2} + \frac{s_2^2}{\bar{x}_2^2} - 2 q_{12} \frac{s_1}{\bar{x}_1} \frac{s_2}{\bar{x}_2} \right\}.$$

7.3 Произведение двух случайных величин

Среднее значение произведения двух случайных величин x_1 и x_2:

$$\bar{y} \approx \bar{x}_1 \bar{x}_2,$$

а ошибка величины $\bar{y}$:

$$s_{\bar{y}}^2 \approx \bar{x}_1^2 \bar{x}_2^2 \left\{ \frac{s_1^2}{\bar{x}_1^2} + \frac{s_2^2}{\bar{x}_2^2} + 2q_{12} \frac{s_1}{\bar{x}_1} \frac{s_2}{\bar{x}_2} \right\}.$$

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.

7.4 Дисперсия произвольной функции от n независимых случайных величин

Еще одним важным частным случаем вычисления матрицы ошибок является вычисление *дисперсии функции* многих независимых случайных переменных.

Если $y = f(x_i)(i = 1, \ldots, n)$, то

$$s_{\bar{y}}^2 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)_{\Big|_{\bar{x}_i}}^2 s^2(\bar{x}_i).$$

7.5 Пример вычисления плотности распределения функции случайных аргументов

Существует еще один способ определить характеристики функции случайных величин $y = y(x_1, x_2, \ldots, x_n)$, но он основан на редко реализуемом в практических задачах предположении, что плотности распределения случайных аргументов этой функции точно известны, [2].

Рассмотрим эту задачу более подробно на примере функции двух случайных аргументов и найдем 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 [2]: если совокупность случайных величин $\{x_1, x_2, \dots x_n\}$ обладает плотностью вероятности $f(x_1, x_2, \dots, x_n)$, то вероятность попадания случайной точки X с координатами $\{x_1, x_2, \dots x_n\}$ в произвольную область G равна интегралу от функции f по этой области:

$$P(X \in G) = \int_G f(u_1, u_2, \dots, u_n)du_1du_2 \dots du_n.$$

Используя эту теорему, получаем:

$$H(t) = \iint\limits_{u+v<t} f_1(u)f_2(v)dudv = \int_{-\infty}^{+\infty} du \int_{-\infty}^{t-u} f_1(u)f_2(v)dv =$$

$$= \int_{-\infty}^{+\infty} du \int_{-\infty}^{t} f_1(u)f_2(w-u)dw,$$

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

8.1 Система линейных уравнений: основные понятия

В общем случае система m линейных уравнений с n неизвестными (или кратко, *линейная система*) (далее — «система») имеет вид [4]; [5]:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \dots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{cases}.$$

Величины $x_1, x_2, \dots, x_n$ — неизвестные, которые нужно вычислить, решив матричное уравнение. Заданная матрица системы (*основная матрица*)

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}.$$

Заданный вектор-столбец свободных членов есть $(b_1, b_2, \dots, b_m)$. 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 $c_1, c_2, \dots, c_n$, которые при подстановке в систему на место неизвестных $x_1, x_2, \dots, x_n$ обращают все уравнения этой системы в тождества.

Не всякая система имеет решение. Например,

$$\begin{cases} x_1 + x_2 = 1 \\ x_1 + x_2 = 2 \end{cases}.$$

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 наличия у системы хотя бы одного решения формулируется *теоремой Кронекера–Капелли*, которая формулируется следующим образом.

Для того чтобы линейная система

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \dots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \end{cases}$$

являлась совместной (имела хотя бы одно решение), необходимо и достаточно, чтобы ранг расширенной матрицы этой системы был равен рангу ее основной матрицы.

Расширенная матрица системы:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} & b_m \end{pmatrix}.$$

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 Решение системы линейных уравнений

8.2.1 Метод Крамера

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. Нужно 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 (основная матрица)

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ -1 & 2 & -3 & 4 \\ 0 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

Распишем подробно вычисление определителя Δ_1 по формуле расчета с помощью алгебраических дополнений.

Напомним, *алгебраическое дополнение* для элемента a_{ij} матрицы A — это число $\Delta_{ij} = (-1)^{i+j} M_{ij}$, где M_{ij} — минор, определитель матрицы, получаемой вычеркиванием i-ой строки и j-го столбца из матрицы A . Для вычисления определителя матрицы A с помощью алгебраических дополнений используется метод разложения по строке (или столбцу) по формуле

$$\Delta = \det A = \sum_{j=1}^{n} a_{ij} \Delta_{ij}.$$

Воспользуемся этим разложением 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

$$x_1 = \Delta_1/\Delta = 1; \;\; x_2 = \Delta_2/\Delta = 2;$$

$$x_3 = \Delta_3/\Delta = 3; \;\; x_4 = \Delta_4/\Delta = 4.$$

8.2.2 Метод Гаусса

Основная идея *метода Гаусса* заключается в том, что со строками и столбцами матрицы можно производить три следующие *элементарные операции*, которые не изменяют ранга матрицы:

- перестановку двух строк (столбцов);
- умножение строки (столбца) на любой отличный от нуля множитель;
- 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:

$$\tilde{A} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \tilde{a}_{22} & \dots & \tilde{a}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \tilde{a}_{m2} & \dots & \tilde{a}_{mn} \end{pmatrix}.$$

5. Все предыдущие шаги осуществить с матрицей

$$\begin{pmatrix} \tilde{a}_{22} & \dots & \tilde{a}_{2n} \\ \vdots & \ddots & \vdots \\ \tilde{a}_{m2} & \dots & \tilde{a}_{mn} \end{pmatrix}.$$

   и т.д. до получения диагональной матрицы.

8.2.3 Замечания о погрешностях матричных операций

Как и метод Крамера, метод Гаусса может привести к достаточно большим погрешностям, если значения коэффициентов и свободных членов заданы приближенно или когда производится округление в процессе вычисления. Это относится, в первую очередь, к случаю, когда основная матрица линейной системы является *плохо обусловленной* (см. раздел 12.6), когда малым изменениям

элементов этой матрицы отвечают достаточно 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, а также итерационные методы Якоби и методы сингулярного матричного разложения [4].

9 Условные и нормальные уравнения

9.1 Понятие о равноточных и неравноточных измерениях

Равноточность результатов (наблюдательных или экспериментальных выборочных данных) означает, что все эти результаты $x_1, x_2, \ldots, x_n$ 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 (среднеквадратическое отклонение):
$$s = \sqrt{\frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

- Средняя квадратичная ошибка среднего арифметического (среднеквадратическое отклонение выбо-

рочного среднего):

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \sqrt{\frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}{n(n-1)}}.$$

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 называются *неравноточными*.

Простейший случай неравноточных измерений — собрание не прямых измерений, а выводов из равноточных измерений, число которых различно в различных выводах. Другими словами, пусть имеется 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 $\{x_1, x_2, \ldots, x_n\}$, и из этого набора нужно вывести наиболее вероятное значение x_k.

Для каждого x_k среднеквадратическое отклонение:

$$s_k = \frac{s}{\sqrt{m_k}},$$

где s — среднеквадратическая ошибка одного измерения.

Если x_k — равноточные, то $s_k = s$. Если x_k — неравноточные, то вместо $s_1, s_2, \ldots, s_n$ вводят числа $p_1, p_2, \ldots, p_n$, называемые *весами измерений*:

$$p_k = \frac{s_0^2}{s_k^2},$$

где s_0^2 — любое положительное число (s_0 называется среднеквадратическим отклонением на единицу веса).

Из определения введенных весов следуют их свойства:

- Веса неравноточных измерений являются обратно пропорциональными 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:

$$\bar{x}_p = \frac{1}{p}\sum_{k=1}^{n} p_k x_k; \;\; p = \sum_{k=1}^{n} p_k.$$

- Наиболее вероятное значение средней квадратичной ошибки измерения с весом единица:

$$s_p = \sqrt{\frac{\sum\limits_{k=1}^{n} p_k (x_k - \bar{x}_p)^2}{n-1}}.$$

- Средняя квадратичная ошибка среднего весового или среднего взвешенного:

$$s_{\bar{x}_p} = \frac{s_0}{\sqrt{p}}; \;\; p = \sum_{k=1}^{n} p_k.$$

9.2 Условные уравнения

В реальных задачах часто бывает так, что подлежащие определению величины нельзя наблюдать непосредственно. Вместо них из наблюдений можно определить только функции неизвестных [12].

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:

$$y_k = \Theta_0 + \Theta_1 x_k + \Theta_2 x_k^2 \quad (k = 1, 2, \ldots, n).$$

В общем виде задача ставится следующим образом: вместо подлежащих определению величин $\Theta_0, \Theta_1, \Theta_2, \ldots$ из наблюдений получаются величины y_k, которые есть функции от неизвестных $\Theta_0, \Theta_1, \Theta_2, \ldots$. Каждое наблюдение дает *условное уравнение* вида

$$f_k(\Theta_0, \Theta_1, \Theta_2, \ldots, x_k) = y_k.$$

Если в процессе наблюдений не было случайных ошибок, или ошибки были так малы, что ими можно было 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, следует взять число наблюдений (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

$\Theta_0^*, \Theta_1^*, \Theta_2^*, \dots$, которые удовлетворяли бы всем условным уравнениям одновременно:

$$f_k(\Theta_0^*, \Theta_1^*, \Theta_2^*, \dots) - y_k = \epsilon_k \neq 0.$$

Величина $f_k(\Theta_0^*, \Theta_1^*, \Theta_2^*, \dots) - y_k$ называется *невязка*.

9.3 Нормальные уравнения

Если дана система равноточных условных уравнений, то следует искать неизвестные таким образом, чтобы сумма квадратов невязок была наименьшей, в чем заключается *принцип Лежандра*.

Образуем сумму квадратов невязок

$$S = \sum_{k=1}^{n} \big[f_k(\Theta_0, \Theta_1, \Theta_2, \dots) - y_k\big]^2.$$

Необходимое условие минимума S:

$$\frac{\partial S}{\partial \Theta_0} = \frac{\partial S}{\partial \Theta_1} = \frac{\partial S}{\partial \Theta_2} = \dots = 0. \tag{14}$$

Полученные уравнения (14) называются *нормальными уравнениями*.

Принцип Лежандра легко обобщается на неравноточные условные уравнения.

Неравноточные уравнения можно привести к равноточным. Пусть известны средние квадратичные ошибки $s_1, s_2, \dots, s_n$ и найдены веса $p_1, p_2, \dots, p_n$. Наиболее вероятная совокупность значений получается при минимизации (*обобщенный принцип Лежандра*)

$$S_p = \sum_{k=1}^{n} p_k \big[f_k(\Theta_0, \Theta_1, \Theta_2, \dots) - y_k\big]^2.$$

Из обобщенного принципа Лежандра легко получить правило приведения неравноточных условных уравнений к равноточным:

$$S_p = \sum_{k=1}^{n} p_k \cdot \varepsilon_k^2 = \sum_{k=1}^{n} \left(\varepsilon_k \sqrt{p_k}\right)^2 = \sum_{k=1}^{n} \tilde{\varepsilon}_k^2.$$

Сделаем важное замечание о работе с условными уравнениями. Составить нормальные уравнения можно при любом виде условных уравнений, но решать их, 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 (вычисление точечных оценок и допустимых интервалов для этих неизвестных).

Гораздо удобнее работать с линейными уравнениями, когда неизвестные зависят от случайных величин линейным образом.

Приводить уравнения к линейному виду можно удачной заменой переменных. Например, пусть условные уравнения имеют вид:

$$\alpha_k \sin(\Theta_0 + \Theta_1) + \beta_k \sin(\Theta_0 - \Theta_1) + \gamma_k \exp\left\{-2\Theta_2\right\} = y_k,$$

где $\Theta_0, \Theta_1, \Theta_2$ — неизвестные (неизвестные параметры исследуемой аналитической модели), которые нужно определить; $\alpha_k, \beta_k, \gamma_k$ — заданные неслучайные величины; y_k — случайные величины.

Сделаем подстановку:

$$\sin(\Theta_0 + \Theta_1) = x;$$

$$\sin(\Theta_0 - \Theta_1) = y;$$

$$\exp\left\{-2\Theta_2\right\} = z.$$

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:

$$\bar{\Theta}_0 = \frac{\arcsin \bar{x} + \arcsin \bar{y}}{2};$$

$$\bar{\Theta}_1 = \frac{\arcsin \bar{x} - \arcsin \bar{y}}{2};$$

$$\bar{\Theta}_2 = -\frac{\ln \bar{z}}{2}.$$

9.4 Общий метод линеаризации условных уравнений

Привести условные уравнения к линейному виду путем замены переменной можно далеко не всегда. Подробно рассмотрим задачу линеаризации условных уравнений с последующим составлением системы нормальных уравнений и ее решением на примере:

$$\Theta_0 \cdot \sin\left(\frac{2\pi t}{\Theta_3} + \Theta_1\right) + \Theta_2 = w.$$

Нужно определить $\Theta_0, \Theta_1, \Theta_2$ и Θ_3, считая, что w содержит случайные ошибки и все измерения равноточные.

Таблица 21

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

Предположим без ограничения общности, что коэффициенты условных уравнений точные, а случайные ошибки малы по модулю.

Случайная величина w задана таблично для разных моментов времени (см. табл. 21).

Подставим в заданный закон табличные значения t_k и w_k и получим систему из 13 нелинейных условных уравнений:

$$\Theta_0 \cdot \sin\left(\frac{2\pi t_k}{\Theta_3} + \Theta_1\right) + \Theta_2 = w_k.$$

9.4.1 Определение начальных условий

Найдем любым способом предварительные приближенные значения параметров $\Theta_0, \Theta_1, \Theta_2, \Theta_3$, опираясь на данные табл. 21.

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, а значит Θ_2 и Θ_0 — 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 неслучайных параметров, положим

$$\Theta_0 = \Theta_0^0 + x_1; \;\; \Theta_1 = \Theta_1^0 + y_1;$$

$$\Theta_2 = \Theta_2^0 + z_1; \;\; \Theta_3 = \Theta_3^0 + u_1.$$

Теперь подлежат определению неизвестные неслучайные параметры: x_1, y_1, z_1, u_1.

9.4.2 Представление условных уравнений в виде ряда по малым параметрам

Подставим

$$\Theta_0 = \Theta_0^0 + x_1; \;\; \Theta_1 = \Theta_1^0 + y_1;$$

$$\Theta_2 = \Theta_2^0 + z_1; \;\; \Theta_3 = \Theta_3^0 + u_1.$$

в условные уравнения, 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

$$f_k = \Theta_0 \cdot \sin\left(\frac{2\pi t_k}{\Theta_3} + \Theta_1\right) + \Theta_2; \;\; y_k = w_k.$$

$$x_1 \cdot \sin\left(\frac{2\pi t_k}{\Theta_3^0} + \Theta_1^0\right) + \Theta_0^0 y_1 \cdot \cos\left(\frac{2\pi t_k}{\Theta_3^0} + \Theta_1^0\right) -$$

$$- u_1 \Theta_0^0 \cdot \frac{2\pi t_k}{\left(\Theta_3^0\right)^2} \cdot \cos\left(\frac{2\pi t_k}{\Theta_3^0} + \Theta_1^0\right) + z_1 +$$

$$+ \left[\Theta_0^0 \cdot \sin\left(\frac{2\pi t_k}{\Theta_3^0} + \Theta_1^0\right) + \Theta_2^0 - w_k\right] = 0;$$

$$\Theta_0^0 = 0{,}90;\ \ \Theta_1^0 = 1{,}50;\ \ \Theta_2^0 = 1{,}08;\ \ \Theta_3^0 = 6{,}30;$$

$$k = 1, 2, \ldots, 13.$$

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; известные неслучайные коэффициенты — a_k, b_k, c_k, d_k. Эти коэффициенты меняются от уравнения к уравнению. Случайные ошибки содержатся только в y_k.

Невязки:

$$\varepsilon_k = a_k x + b_k y + c_k z + d_k u + y_k.$$

9.4.3 Получение системы нормальных уравнений для первого приближения

Согласно принципу Лежандра, нужно минимизировать

$$S = \sum_{k=1}^{n} \varepsilon_k^2 = \sum_{k=1}^{n} (a_k x + b_k y + c_k z + d_k u + y_k)^2.$$

Считая условные уравнения равноточными, запишем необходимое условие минимума:

$$\begin{cases} \dfrac{\partial S}{\partial x} = 2 \sum\limits_{k=1}^{n} \left(a_k x + b_k y + c_k z + d_k u + y_k\right) \cdot a_k = 0 \\ \dfrac{\partial S}{\partial y} = 2 \sum\limits_{k=1}^{n} \left(a_k x + b_k y + c_k z + d_k u + y_k\right) \cdot b_k = 0 \\ \dfrac{\partial S}{\partial z} = 2 \sum\limits_{k=1}^{n} \left(a_k x + b_k y + c_k z + d_k u + y_k\right) \cdot c_k = 0 \\ \dfrac{\partial S}{\partial u} = 2 \sum\limits_{k=1}^{n} \left(a_k x + b_k y + c_k z + d_k u + y_k\right) \cdot d_k = 0 \end{cases}.$$

Эти условия приводят к нормальным уравнениям

$$\begin{cases} x \sum a_k^2 + y \sum a_k b_k + z \sum a_k c_k + u \sum a_k d_k + \sum a_k y_k = 0 \\ x \sum b_k a_k + y \sum b_k^2 + z \sum b_k c_k + u \sum b_k d_k + \sum b_k y_k = 0 \\ x \sum c_k a_k + y \sum c_k b_k + z \sum c_k^2 + u \sum c_k d_k + \sum c_k y_k = 0 \\ x \sum d_k a_k + y \sum d_k b_k + z \sum d_k c_k + u \sum d_k^2 + \sum d_k y_k = 0 \end{cases}.$$

9.4.4 Решение системы нормальных уравнений для первого приближения

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

$$\bar{x} = \frac{\Delta_x}{\Delta} = -\sum_{k=1}^{n} \frac{\Delta_k}{\Delta} y_k;$$

$$s_{\bar{x}}^2 = \sum_{k=1}^{n} \frac{\Delta_k^2}{\Delta^2} s_k^2. \qquad (15)$$

Докажем утверждение (15).

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} — 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:

$$\sum b_k\Delta_k = \sum c_k\Delta_k = \sum d_k\Delta_k = 0;$$

$$\sum a_k\Delta_k = \Delta.$$

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

$$s_{\bar{x}}^2 = \frac{\Delta_{11}}{\Delta} \cdot s_0^2; \;\; p_{\bar{x}} = \frac{s_0^2}{s_{\bar{x}}^2} = \frac{\Delta}{\Delta_{11}}.$$

Аналогично для других неизвестных:

$$\bar{y} = \frac{\Delta_y}{D}; \;\; \bar{z} = \frac{\Delta_z}{\Delta}; \;\; \bar{u} = \frac{\Delta_u}{\Delta};$$

$$p_{\bar{y}} = \frac{\Delta}{\Delta_{22}}; \;\; p_{\bar{z}} = \frac{\Delta}{\Delta_{33}}; \;\; p_{\bar{u}} = \frac{\Delta}{\Delta_{44}},$$

где Δ — определитель основной матрицы линейной системы нормальных уравнений; 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

$$\bar{S} = \sum_{k=1}^{n} \varepsilon_k^2 = \sum_{k=1}^{n} \big(a_k\bar{x} + b_k\bar{y} + c_k\bar{z} + d_k\bar{u} + y_k\big)^2 =$$

$$= \bar{x}\sum_{k=1}^{n} a_k y_k + \bar{y}\sum_{k=1}^{n} b_k y_k + \bar{z}\sum_{k=1}^{n} c_k y_k + \bar{u}\sum_{k=1}^{n} d_k y_k + \sum_{k=1}^{n} y_k^2.$$

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

$$s_{\bar{x}} = \frac{s_0}{\sqrt{p_{\bar{x}}}};\quad s_{\bar{y}} = \frac{s_0}{\sqrt{p_{\bar{y}}}};$$

$$s_{\bar{z}} = \frac{s_0}{\sqrt{p_{\bar{z}}}};\quad s_{\bar{u}} = \frac{s_0}{\sqrt{p_{\bar{u}}}}.$$

Решение задачи (в первом приближении) записывается как

$$x = \bar{x} \pm s_{\bar{x}};\quad y = \bar{y} \pm s_{\bar{y}};\quad z = \bar{z} \pm s_{\bar{z}};\quad u = \bar{u} \pm s_{\bar{u}}.$$

Для рассматриваемого случая:

$$\bar{x} = 0{,}096;\quad \bar{y} = 0{,}092;\quad \bar{z} = -0{,}075;\quad \bar{u} = -0{,}026;$$

$$p_{\bar{x}} = 6{,}2;\quad p_{\bar{y}} = 0{,}40;\quad p_{\bar{z}} = 4{,}5;\quad p_{\bar{u}} = 0{,}113;$$

$$\bar{S} = 0{,}006925;$$

$$s_0^2 = \frac{0{,}007}{13-4} = 0{,}000778;\quad s_0 = 0{,}028;$$

$$s_{\bar{x}}^2 = \frac{0{,}000778}{6{,}2} = 0{,}00013;\quad s_{\bar{x}} = 0{,}011;$$

$$s_{\bar{y}}^2 = \frac{0{,}000778}{0{,}40} = 0{,}00019;\quad s_{\bar{y}} = 0{,}014;$$

$$s_{\bar{z}}^2 = \frac{0{,}000778}{4{,}5} = 0{,}00017;\quad s_{\bar{z}} = 0{,}013;$$

$$s_{\bar{u}}^2 = \frac{0{,}000778}{0{,}113} = 0{,}0069;\quad s_{\bar{u}} = 0{,}083.$$

Окончательно решение первого приближения:

$$x = 0{,}096 \pm 0{,}011; y = 0{,}092 \pm 0{,}014;$$

$$z = -0{,}075 \pm 0{,}013; u = -0{,}026 \pm 0{,}083.$$

Таким образом, решение исходной системы n условных уравнений

$$\Theta_0 \cdot \sin\left(\frac{2\pi t}{\Theta_3} + \Theta_1\right) + \Theta_2 = w,$$

где w и t — дискретные наборы n экспериментальных данных, в первом приближении:

$$\Theta_0 = 0{,}90 + 0{,}096 \pm 0{,}011;$$

$$\Theta_1 = 1{,}50 + 0{,}092 \pm 0{,}014;$$

$$\Theta_2 = 1{,}08 - 0{,}075 \pm 0{,}013;$$

$$\Theta_3 = 6{,}30 - 0{,}026 \pm 0{,}083.$$

9.4.5 Стратегия дальнейшего решения

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 последние.

Процесс может быть остановлен тогда, когда в двух последовательных итерациях с заданной точностью получают одинаковые значения. Так, точечные оценки решения второго приближения:

$$\{\bar{x}, \bar{y}, \bar{z}, \bar{u}\} = \{0{,}004; -0{,}013; -0{,}001; -0{,}006\};$$

$$\bar{S} = 0{,}006864.$$

10 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 (силу) влияния внешних факторов на статистическую совокупность выборочных данных или влияние двух выборок друг на друга можно определить методами *корреляционного анализа*. Конкретная математическая модель влияния устанавливается *регрессионным анализом* [6].

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 позволяет 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 будут рассмотрены в разделе 13). Кроме того,

надо проверить, чтобы дисперсии выборок были одинаковыми (проверка равенства дисперсий двух выборок производится по F-критерию Фишера).

Рассмотрим влияние фактора $A = \{A_1, A_2, \ldots, A_k\}$. В каждом столбце (см. табл. 22) выборка $\{x_{i1}, x_{i2}, \ldots, x_{in}\}$ характеризует изменение данных под влиянием фактора A уровня A_i. Точечная оценка дисперсии такой i-й 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 $i = 1, 2, \ldots, k$ все значения $s_i^2 = \text{const}$. 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 объединенной выборки:

$$\bar{x} = \frac{1}{n \cdot k} \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij}.$$

Дисперсия объединенной выборки:

$$s^2 = \frac{1}{n \cdot k - 1} \sum_{i=1}^{k} \sum_{j=1}^{n} \left(x_{ij} - \bar{x} \right)^2 =$$

$$= \frac{1}{n \cdot k - 1} \sum_{i=1}^{k} \sum_{j=1}^{n} \left(x_{ij} - \frac{1}{n \cdot k} \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij} \right)^2 =$$

$$= \frac{1}{n \cdot k - 1} \left[(n \cdot k - k) \cdot s_0^2 + (k-1) \cdot s_A^2 \right].$$

Таблица 22

Уровни фактора $A = \{A_1, A_2, \ldots, A_i, \ldots, A_k\}$ при однофакторном дисперсионном анализе.

Номер наблюдения	A_1	A_2	$\ldots$	A_i	$\ldots$	A_k
1	x_{11}	x_{21}	$\ldots$	x_{i1}	$\ldots$	x_{k1}
2	x_{12}	x_{22}	$\ldots$	x_{i2}	$\ldots$	x_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	x_{1j}	x_{2j}	$\ldots$	x_{ij}	$\ldots$	x_{kj}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{1n}	x_{2n}	$\ldots$	$\vdots$	$\ldots$	x_{kn}

Суммарная дисперсия (рассеяние внутри выборки):

$$s_0^2 = \frac{1}{n \cdot k - k} \sum_{i=1}^{k} (n-1) \cdot s_i^2.$$

Дисперсия между выборками (сумма квадратов отклонений средних по выборке от среднего объединенной выборки):

$$s_A^2 = \frac{1}{k-1} \sum_{i=1}^{k} n \cdot \left(\frac{1}{n} \sum_{j=1}^{n} x_{ij} - \frac{1}{n \cdot k} \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij} \right)^2.$$

Докажем тождество, представляющее собой важную связь дисперсии объединенной выборки с дисперсией внутри одной выборки и с дисперсией между выборками:

$$s^2 = \frac{1}{n \cdot k - 1} \left[(n \cdot k - k) \cdot s_0^2 + (k-1) \cdot s_A^2 \right].$$

Для доказательства и в методических целях распишем выражения s^2, s_0^2, s_A^2:

$$s^2 \cdot (n \cdot k - 1) = \sum_{i=1}^{k} \sum_{j=1}^{n} \left(x_{ij} - \frac{1}{n \cdot k} \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij} \right)^2 =$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n} \left\{ x_{ij}^2 - \frac{2}{n \cdot k} x_{ij} \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij} + \frac{1}{(n \cdot k)^2} \left(\sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij} \right)^2 \right\} =$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij}^2 - \frac{2}{n \cdot k} \left(\sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij} \right)^2 + \frac{n \cdot k}{(n \cdot k)^2} \left(\sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij} \right)^2 =$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij}^2 - \frac{1}{n \cdot k} \left(\sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij} \right)^2;$$

$$s_0 = \frac{1}{n \cdot k - k} \sum_{i=1}^{k} (n-1) \cdot \frac{1}{n-1} \sum_{j=1}^{n} \left(x_{ij} - \frac{1}{n} \sum_{j=1}^{n} x_{ij} \right)^2 =$$

$$= \frac{1}{k \cdot (n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} \left(x_{ij} - \frac{1}{n} \sum_{j=1}^{n} x_{ij} \right)^2 =$$

$$= \frac{1}{k \cdot (n-1)} \left\{ \sum_{i=1}^{k} \sum_{j=1}^{n} \left[x_{ij}^2 - \frac{2}{n} x_{ij} \sum_{j=1}^{n} x_{ij} + \frac{1}{n^2} \left(\sum_{j=1}^{n} x_{ij} \right)^2 \right] \right\} =$$

$$= \frac{1}{k \cdot (n-1)} \left\{ \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij}^2 - \frac{2}{n} \sum_{i=1}^{k} \sum_{j=1}^{n} \left(x_{ij} \sum_{j=1}^{n} x_{ij} \right) + \right.$$

$$\left. + \frac{1}{n^2} \sum_{i=1}^{k} \sum_{j=1}^{n} \left(\sum_{j=1}^{n} x_{ij} \right)^2 \right\} =$$

$$= \frac{1}{k \cdot (n-1)} \left\{ \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij}^2 - \frac{1}{n} \sum_{i=1}^{k} \left(\sum_{j=1}^{n} x_{ij} \right)^2 \right\};$$

$$s_A = \frac{1}{k-1}\sum_{i=1}^{k} n \cdot \left(\frac{1}{n}\sum_{j=1}^{n} x_{ij} - \frac{1}{n\cdot k}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2 =$$

$$= \frac{1}{n\cdot(k-1)}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ij} - \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2 =$$

$$= \frac{1}{n\cdot(k-1)}\left\{\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ij}\right)^2 - \frac{2}{k}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij} + \right.$$

$$\left. + \sum_{i=1}^{k}\frac{1}{k^2}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2\right\} =$$

$$= \frac{1}{n\cdot(k-1)}\left\{\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ij}\right)^2 - \frac{1}{k}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2\right\}.$$

Таким образом, нужно доказать, что

$$\frac{1}{n\cdot k-1}\cdot\left[\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \frac{1}{n\cdot k}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2\right] =$$

$$= \frac{1}{n\cdot k-1}\cdot\left[(n\cdot k-k)\cdot\left[\frac{1}{k\cdot(n-1)}\left\{\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \right.\right.\right.$$

$$\left.\left.\left. - \frac{1}{n}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ij}\right)^2\right\}\right] +$$

$$+(k-1)\cdot\left[\frac{1}{n\cdot(k-1)}\left\{\sum_{i=1}^{k}\left(\sum_{j=1}^{n} x_{ij}\right)^2 - \frac{1}{k}\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2\right\}\right]\right].$$

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.

Отметим также, что величину $s^2 \cdot (nk-1)$ называют *общей суммой*, величину $s_A^2 \cdot (k-1)$ — *факторной суммой*, а их разность — *остаточной суммой*.

Проверка влияния фактора A на изменение средних сводится к сравнению 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 γ, если

$$\frac{s_A^2}{s_0^2} > F_\gamma(k-1, k(n-1)),$$

где $F_\gamma(k-1, k(n-1))$ есть γ-квантиль[15] $F_\gamma(k_1, k_2)$ распределения Фишера с $k_1 = (k-1), k_2 = k(n-1)$ степенями свободы.

[15] Или $\alpha = (1-\gamma)$-процентная точка.

11 Корреляционный анализ

Корреляционный анализ предполагает изучение зависимости между случайными величинами с одновременной количественной оценкой степени неслучайности их совместного изменения [6]. Зависимость между случайными величинами X и Y характеризуется *коэффициентом корреляции*, точное значение которого

$$\rho = \frac{M[(X - m_x) \cdot (Y - m_y)]}{\sqrt{D[X] \cdot D[Y]}}.$$

Здесь m_x — математическое ожидание X; m_y — математическое ожидание Y.

Коэффициент корреляции был введен в разделе 4.6. Он показывает, насколько зависимость между случайными величинами X и Y близка к строго линейной. Если X и Y имеют нормальное распределение, то $\rho = 0$ для них означает отсутствие линейной связи (отсутствие *линейной корреляции*), хотя при этом они могут быть зависимы. Равенство $\rho = \pm 1$ означает наличие строгой линейной связи.

Для колличественной оценки нелинейной связи используют понятие *криволинейной корреляции.*

11.1 Оценка коэффициента корреляции

Как было приведено в разделе 4.6, в случае работы с реальными данными для двух случайных величин $X = \{x_1, x_2, \ldots, x_n\}$ и $Y = \{y_1, y_2, \ldots, y_n\}$ *выборочный коэффициент корреляции*, обозначаемый в литературе по-

разному, $r = q = \rho^*$:

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

где

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

При малом объеме выборки $n < 15$, коэффициент корреляции лучше оценивать по формуле [6]:

$$\tilde{r} = r \cdot \left[1 + \frac{1 - r^2}{2(n-3)}\right].$$

При большом объеме выборки $n > 200$ выборочный коэффициент корреляции r имеет нормальное распределение:

$$r \sim N(m_r, \sigma_r^2);$$

$$m_r = r; \;\; \sigma_r^2 = \frac{1 - r^2}{n-1}.$$

11.2 Исследование значимости корреляции

На практике особенно важно исследование значимости корреляции, насколько сильно коэффициент корреляции отличен от нуля. Для этой цели вычисляется выборочное значение коэффициента корреляции r и сравнивается с табличным критическим значением r_γ (см. табл. 23).

Для выборки большого объема $n > 200$ критическое значение коэффициента корреляции хорошо аппрокси-

Таблица 23

Некоторые критические значения r_γ выборочного коэффициента корреляции.

Количество элементов выборки n	$\gamma = 0{,}90$	$\gamma = 0{,}95$	$\gamma = 0{,}99$
3	0,998	0,997	1,000
10	0,549	0,632	0,765
15	0,441	0,514	0,641
20	0,378	0,444	0,561

Таблица 24

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

мируется u_γ-квантилем нормального распределения (см. табл. 14):

$$r_\gamma = \frac{1}{\sqrt{n-1}} \cdot u_\gamma.$$

ПРИМЕР. Определим значимость корреляционной зависимости между двумя заданными выборками [6].

В результате наблюдений над случайными величинами X и Y получена совокупность данных из 10 элементов для каждой случайной величины (см. табл. 24).

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:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{723}{\sqrt{405{,}6 \cdot 1422}} = 0{,}952.$$

С помощью приближенной оценки, лучшей для малых выборок, можно оценить коэффициент корреляции следующим образом:

$$\tilde{r} = r \cdot \left[1 + \frac{1 - r^2}{2(n-3)}\right] = 0{,}952 \cdot \left(1 + \frac{1 - 0{,}952^2}{2 \cdot 7}\right) = 0{,}958.$$

Теперь, для сравнения, используя u_γ-квантиль нормального распределения (см. табл. 14), применим формулу для оценки критического значения коэффициента корреляции для больших выборок:

$$r_\gamma = r_{0{,}95} = \frac{1}{\sqrt{n-1}} \cdot u_\gamma = \frac{1{,}96}{3} = 0{,}653.$$

Учитывая малость выборки ($n = 10$), следует использовать величину

$$r_{0{,}95} = 0{,}632,$$

которую можно получить из таблицы точных критических значений r_γ (см. табл. 23 [6]).

В любом случае $\tilde{r} = 0{,}958 > 0{,}653$, и корреляция признается значимой с надежностью $\gamma = 0{,}95$.

11.3 Понятие криволинейной корреляции

Для количественной оценки нелинейной связи между двумя случайными величинами X и Y вводят новые характеристики, меняющиеся от 0 до 1:

- Неслучайная величина η_{yx} — выборочное корреляционное отношение Y к X;
- Неслучайная величина η_{xy} — выборочное корреляционное отношение X к Y.

Рассмотрим, для определенности, величину

$$\eta = \eta_{yx} = \sigma_{\bar{y}_x}/\sigma_y,$$

где

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

есть *межгрупповое среднеквадратическое отклонение* и *общее среднеквадратическое отклонение*, соответственно.

Величина $n = n_x + n_y$; $\bar{y}_x$ — условное среднее случайной величины Y; y — элемент выборки, и суммы записываются по всем элементам. Количество элементов для X и Y есть n_x и n_y соответственно.

Если $\eta = 0$, то корреляционной зависимости нет. Действительно, в этом случае межгрупповая дисперсия равна нулю, а значит при всех значениях X условные средние сохраняют постоянное значение. Другими словами, условное среднее не зависит от X.

Если $\eta = 1$, то имеется функциональная зависимость. Например *параболическая корреляция второго порядка*

$$\bar{y}_x = ax^2 + bx + c$$

или *параболическая корреляция третьего порядка*

$$\bar{y}_x = ax^3 + bx^3 + cx + d.$$

Соответствующие параметры a, b, c, d находятся методами полиномиального регрессионного анализа, который рассматривается в разделе 12.6.

Наконец, если выборочное корреляционное отношение равно модулю коэффициента линейной корреляции, которое необходимо предварительно вычислить, то имеет место точная линейная корреляция, параметры которой находятся методами линейного регрессионного анализа, рассматриваемого в разделе 12.1.

12 Регрессионный анализ

Методы дисперсионного и корреляционного анализа позволяют определить, есть ли связь между случайными величинами, и оценить силу этой связи. Необходимо уметь определять и конкретный вид функциональной зависимости между случайными величинами — в этом заключается задача регрессионного анализа.

Пусть исследуется связь между двумя случайными выборками

$$X = \{x_1, x_2, \ldots, x_n\};\ \ Y = \{y_1, y_2, \ldots, y_n\}.$$

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 МНК (см. раздел 6.2.3). Согласно этому методу, в качестве уравнения регрессии $y = f(x)$ выбирается функция, которая дает минимум сумме квадратов разностей

$$S = \sum_{i=1}^{n} \Big[y_i - f(x_i) \Big]^2.$$

На практике вид функции $f(x)$ заранее известен, она представляет собой полином некоторой степени, и путем минимизации находятся коэффициенты этого полинома.

Отметим, что принцип минимизации суммы квадратов разностей эффективен в том случае, если эти разности имеют нормальное распределение. Однако если

$\varepsilon_i = y_i - f(x_i)$ подчиняются другому закону распределения, то следует минимизировать не S. Так, если разности независимы и имеют двустороннее экспоненциальное распределение, заданное плотностью

$$f(\varepsilon_i) = \frac{1}{2\sigma} \cdot \exp\left\{ - \frac{|\varepsilon_i|}{\sigma} \right\}$$

(сравните с плотностью нормального распределения

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

то следует минимизировать сумму модулей:

$$\tilde{S} = \sum_{i=1}^{n} \Big| y_i - f(x_i) \Big|.$$

В дальнейшем будем рассматривать только нормально распределенные разности.

Количественная мера рассеяния значений y_i вокруг регрессии $f(x)$ — 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 (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*:

$$f(x) = a_0 + a_1 \cdot x + a_2 \cdot x^2 + \ldots.$$

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 Чебышёва, см. раздел 12.6.3).

Общая (наиболее краткая) 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

$$\{x_1, x_2, \ldots, x_n\};\ \ \{y_1, y_2, \ldots, y_n\},$$

т.е. нахождение неизвестных коэффициентов a и b методом МНК из условия минимума суммы квадратов разностей

$$\sum_{i=1}^{n} \big[y_i - f(x_i)\big]^2.$$

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.

12.1 Постановка задачи линейного регрессионного анализа

Модель зависимости двух случайных выборок в линейном регрессионном анализе (система линейных условных уравнений, см. раздел 9):

$$y = f(x) = \alpha + \beta \cdot x,$$

где α и β — истинные, никогда неизвестные, коэффициенты регрессии. Их искомые выборочные оценки будем обозначать a и b соответственно.

Условие минимума по α и β суммы квадратов разностей:

$$S = \sum_{i=1}^{n} \big[y_i - f(x_i)\big]^2$$

дает систему двух уравнений (система линейных нормальных уравнений, см. раздел 9):

$$\begin{cases} \dfrac{\partial S}{\partial \alpha} = \displaystyle\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \big(\alpha + \beta \cdot x_i\big) = 0 \\ \\ \dfrac{\partial S}{\partial \beta} = \displaystyle\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \big(\alpha + \beta \cdot x_i\big) \cdot x_i = 0, \end{cases}$$

из которой следует система

$$\begin{cases} n \cdot \alpha + \beta \displaystyle\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \\ \alpha \displaystyle\sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i. \end{cases}$$

Решение системы:

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2};$$

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}.$$

Для проверки правильности вычислений можно использовать соотношение:

$$\bar{y} = a + b \cdot \bar{x},$$

где

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

Важной особенностью регрессионных уравнений является тот факт, что 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

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2; \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2,$$

то регрессии $y = f(x)$ и $x = \varphi(y)$ можно записать следующим образом:

$$y = \bar{y} + r \cdot \frac{s_y}{s_x} \cdot \left(x - \bar{x}\right);$$

$$x = \bar{x} + r \cdot \frac{s_x}{s_y} \cdot \left(y - \bar{y}\right),$$

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.

12.2 Статистический анализ параметров линейной регрессии

Рассмотрим статистический анализ найденных оценок коэффициентов a и b 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 в модели $y = f(x) = \alpha + \beta \cdot x$ означает, что модель линейной регрессии не подходит.

Для проверки значимости отклонения β от нуля используется статистика t-распределения Стьюдента. Истинное значение β считается значимо отличающимся от нуля с доверительной вероятностью γ (процентной точкой $\alpha = 1 - \gamma$), если найденная оценка b:

$$|b| > t_{n-2,\gamma} \cdot s_\beta = T^{-1}\Big(n-2, \frac{1+\gamma}{2}\Big) \cdot s_\beta.$$

Интервал с доверительной вероятностью γ для истинного коэффициента β определяется как:

$$b - s_\beta \cdot T^{-1}\Big(n-2, \frac{1+\gamma}{2}\Big) \leqslant \beta \leqslant b + s_\beta \cdot T^{-1}\Big(n-2, \frac{1+\gamma}{2}\Big),$$

где

$$b = \frac{n\sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} x_i^2 - \big(\sum\limits_{i=1}^{n} x_i\big)^2}.$$

Здесь число степеней свободы t-распределения есть $n-2$, поскольку оцениваются два неизвестных коэффициента. Значения функции T^{-1} см. табл. 15. Далее,

$$s_\beta = \frac{s}{s_x \cdot \sqrt{n-1}}; \tag{16}$$

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} \big(y_i - a - bx_i\big)^2.$$

Другая форма записи последнего выражения:

$$s^2 = \frac{1}{n-2} \cdot S,$$

где S — сумма квадратов невязок. В выражении (16) величина s_x^2 есть

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

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

Аналогично предыдущему интервал с доверительной вероятностью γ для истинного коэффициента α (не путайте обозначение коэффициента и обозначение процентной точки) определяется как

$$a - s_\alpha \cdot T^{-1}\left(n-2, \frac{1+\gamma}{2}\right) \leqslant \alpha \leqslant a + s_\alpha \cdot T^{-1}\left(n-2, \frac{1+\gamma}{2}\right),$$

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 нормированная разность вычисленного и истинного коэффициентов регрессии имела известное табличное распределение (распределение Стьюдента):

$$\frac{b-\beta}{s_\beta} \sim t_{n-2,\gamma}; \quad \frac{a-\alpha}{s_\alpha} \sim t_{n-2,\gamma}.$$

Вывод 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 регрессии в числителе формируется величина, обладающая нормальным законом распределения, а в знаменателе формируется величина, обладающая распределением $\sqrt{\chi^2_\gamma(n-2)}$ [10].

ПРИМЕР. Пусть задана совокупность данных (см. табл. 25). Требуется оценить параметры линейной регрессии. Для этих данных нужно найти точечную и интервальную оценки коэффициентов α и β линейной регрессии $y = \alpha + \beta \cdot x$. Принять доверительную вероятность $\gamma = 0{,}95$.

Таблица 25

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

В этом примере введем для компактности обозначений $\sum_{i=1}^{n} = \sum_{i=1}^{10} = \sum$.

Сначала вычислим точечные оценки β и α. Для

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

вычислим необходимые суммы[16]:

$$\sum x_i = 61{,}7;\ \left(\sum x_i\right)^2 = 3806{,}89;\ \sum x_i^2 = 486{,}99;$$

$$\sum y_i = 340;\ \sum x_i y_i = 2695{,}1.$$

Тогда

$$b = \frac{10 \cdot 2695{,}1 - 61{,}7 \cdot 340}{10 \cdot 486{,}99 - 3806{,}89} = 5{,}6189;$$

$$a = \frac{\sum y_i - b\sum x_i}{n} = \frac{340 - 5{,}6189 \cdot 61{,}7}{10} = -0{,}669.$$

Построим доверительные интервалы для α и β. Предварительно вычислим:

$$\bar{x} = 6{,}17;\ s_x^2 = \frac{1}{n-1}\sum \left(x_i - \bar{x}\right)^2 = 11{,}8112;\ s_x = 3{,}4367.$$

[16] При обработке реальных данных перед началом вычислений рекомендуется написать соответствующие простые программы.

Далее вычислим дисперсию

$$s^2 = \frac{1}{n-2} \sum \left(y_i - \hat{y}_i\right)^2; \;\; \hat{y}_i = a + b \cdot x_i.$$

Для нашей задачи

$$\hat{y}_i = \{6{,}074; 12{,}816; 15{,}064; 22{,}930; 32{,}483; 37{,}540;$$

$$44{,}844; 51{,}025; 56{,}082; 61{,}139\}.$$

Тогда

$$s^2 = \frac{1}{8} \sum \left(y_i - \hat{y}_i\right)^2 = 13{,}4755;$$

$$s_\beta = \frac{s}{s_x \cdot \sqrt{n-1}} = 0{,}3560;$$

$$s_\alpha = s \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1) \cdot s_x^2}} = 2{,}485.$$

Для $\gamma = 0{,}95$ имеем (см. табл. 15)

$$t_{10-2;0,95} =$$

$$= T^{-1}\left(10-2; \frac{1+0{,}95}{2}\right) = T^{-1}\left(8; 0{,}975\right) = 2{,}306.$$

Если искомое линейное приближение верно, то коэффициент b должен быть, очевидно, значимо отличен от нуля. Это проверяется также t-статистикой:

$$|b| = 5{,}6189 > T^{-1}\left(8; 0{,}975\right) \cdot s_\beta = 2{,}306 \cdot 0{,}356 = 0{,}821.$$

Можно проверить, есть ли основания округлить коэффициент β до 5.

$$|5{,}619 - 5{,}0| = 0{,}619 < T^{-1}\left(8; 0{,}975\right) \cdot s_\beta = 0{,}821.$$

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:

$$5{,}619 - 2{,}306 \cdot 0{,}356 \leqslant \beta \leqslant 5{,}619 + 2{,}306 \cdot 0{,}356;$$

$$4{,}800 \leqslant \beta \leqslant 6{,}462.$$

Аналогично исследуем оценки для коэффициента α. Сначала проверим, можно ли с хорошей точностью считать этот коэффициент равным нулю:

$$|a| = 0{,}669 < T^{-1}\big(8; 0{,}975\big) \cdot s_\alpha = 2{,}306 \cdot 2{,}485 = 5{,}73.$$

Неравенство выполняется, следовательно, на заданном уровне точности α неотличима от нуля. Двусторонний доверительный интервал для α:

$$-0{,}669 - 2{,}306 \cdot 2{,}485 \leqslant \alpha \leqslant -0{,}669 + 2{,}306 \cdot 2{,}485;$$

$$-6{,}399 \leqslant \alpha \leqslant 5{,}062.$$

Окончательно получаем, что с доверительной вероятностью $\gamma = 0{,}95$ уравнение искомой линейной регрессии:

$$y = 5 \cdot x.$$

12.3 Коэффициент детерминации

Здесь и далее в этом разделе введем для компактности обозначений $\sum_{i=1}^{n} = \sum$.

При первичном анализе данных одной из очень полезных характеристик точности подбора регрессии является R^2–статистика, в которой вычисляется *коэффициент детерминации*

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}.$$

Этот коэффициент показывает меру вклада регрессии в общее отклонение от среднего и выражает корреляцию между y и его регрессией $\hat{y}$.

Для линейной регрессии величина $(R^2 \cdot 100\%)$ показывает, сколько процентов общего отклонения от среднего объясняется самим уравнением регрессии $\hat{y} = a + bx$. Формула для расчета коэффициента детерминации

$$R^2 = \frac{\Big(\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i\Big)^2}{\Big(\sum x_i^2 - \frac{1}{n}\big(\sum x_i\big)^2\Big) \cdot \Big(\sum y_i^2 - \frac{1}{n}\big(\sum y_i\big)^2\Big)}.$$

Кроме того, этот коэффициент можно вычислить следующим образом:

$$R^2 = \frac{\nu_1 \tilde{F}}{\nu_1 F + \nu_2},$$

где $\nu_1 = 1; \nu_2 = n - l$ (l — число параметров; для линейной регрессии $l = 2$); $\tilde{F}$ — расчетная статистика распределения Фишера:

$$\tilde{F} = \Big(\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i\Big) \cdot (n-2) \cdot \Bigg[\Big(\sum x_i^2 - \\ - \frac{1}{n}\big(\sum x_i\big)^2\Big) \cdot \Big(\sum y_i^2 - \frac{1}{n}\big(\sum y_i\big)^2\Big) - \Big(\sum x_i y_i - \\ - \frac{1}{n}\sum x_i \sum y_i\Big)^2\Bigg]^{-1}$$

и сравнивается с табличной

$$F(\nu_1, \nu_2) = F(1, n-2) = (t_{n-2})^2.$$

Так, если

$$\tilde{F} > 4 \cdot F(1, n-2),$$

то линейная регрессия может описывать данные. Далее, вычисляя коэффициент детерминации R^2, определяем количество процентов общего отклонения от среднего, которое объясняется регрессией.

Для данных табл. 24: $R^2 = 90{,}6\%$.

Для данных табл. 25: $R^2 = 96{,}8\%$.

12.4 Анализ остатков

Кратко рассмотрим вопрос об анализе остатков, получаемых по регрессионному уравнению

$$e_i = y_i - \hat{y}_i; \;\; \sum e_i = 0;$$

(введено для компактности обозначений $\sum_{i=1}^{n} = \sum$).

Остатки содержат информацию о том, почему построенная модель может противоречить наблюдениям.

Остаток можно представить в виде суммы случайной компоненты остатков и систематической компоненты остатков:

$$e_i = q_i + B_i,$$

причем

$$q_i = (y_i - \hat{y}_i) - (M[y_i] - M[\hat{y}_i]); \;\; B_i = M[y_i] - M[\hat{y}_i].$$

Регрессионная модель *корректна*, если $B_i = 0$. Для корректной модели остатки есть наблюдаемые ошибки, которые являются независимыми, обладают нулевым средним, одинаковыми дисперсиями и подчиняются нормальному распределению.

Вообще говоря, при обработке реальных данных между остатками существует зависимость. Укажем дальнейшие пути исследования остатков:

- исследование на нормальность распределения;

- исследование на постоянство дисперсии и в случае обнаружения непостоянства использование в модели взвешенного МНК;
- исследование влияния времени;
- исследование зависимости от степени регрессионного полинома (в случае полиномиальной регрессии).

12.5 Оценка остаточной дисперсии и сравнение двух линейных регрессий

Рассмотрим задачу сравнения двух линейных регрессий и анализ остаточных дисперсий на примере.

ПРИМЕР. Пусть в ходе двух независимых экспериментов получены результаты ($n_1 = 10, n_2 = 6$), указанные в табл. 26-27. Проверим, являются ли статистически неразличимыми линейные регрессионные модели, полученные по обеим выборкам. Доверительная вероятность $\gamma = 0{,}95$.

Первое, что нужно сделать — это выписать регрессионные модели для обеих выборок.

Как и раньше, для удобства обозначим для первой и второй выборок

$$\sum_{i=1}^{n_1} = \sum_{i=1}^{10} = \sum_{1}; \quad \sum_{i=1}^{n_2} = \sum_{i=1}^{6} = \sum_{2}.$$

Тогда для первой выборки:

$$\sum_{1} x_{1i} = 141; \quad \bar{x}_{1i} = 14{,}1;$$

$$\Big(\sum_{1} x_{1i}\Big)^2 = 19881; \quad \sum_{1} x_{1i}^2 = 2789;$$

Таблица 26

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 27

Статистический ряд распределения x_{2i}, y_{2i}.

x_{2i}	12	16	21	23	28	31
y_{2i}	54	68	87	93	112	130

$$\sum_1 y_{1i} = 579; \quad \bar{y}_{1i} = 57{,}9; \quad \sum_1 x_{1i} y_{1i} = 11361.$$

Коэффициенты регрессии для первой выборки:

$$b_1 = \frac{n_1 \sum\limits_1 x_i y_i - \sum\limits_1 x_i \sum\limits_1 y_i}{n_1 \sum\limits_1 x_i^2 - \Big(\sum\limits_1 x_i\Big)^2} =$$

$$= \frac{10 \cdot 11361 - 141 \cdot 579}{10 \cdot 2789 - 19881} = 3{,}992;$$

$$a_1 = \frac{\sum\limits_1 y_i - b_1 \sum\limits_1 x_i}{n_1} = \frac{579 - 3{,}992 \cdot 141}{10} = 1{,}613.$$

Кроме того, далее понадобится выборочная дисперсия для x_{1i}:

$$s^2_{\bar{x}_1} = \frac{1}{n_1 - 1} \sum_1 \Big(x_{1i} - \bar{x}_1\Big)^2 = 88{,}989;$$

$$s_{\bar{x}_1} = 9{,}433.$$

Аналогичные вычисления проведем для второй выборки:

$$\sum_2 x_{2i} = 131; \;\; \bar{x}_{2i} = 21{,}83;$$

$$\Big(\sum_2 x_{2i}\Big)^2 = 17161; \;\; \sum_2 x_{2i}^2 = 3115;$$

$$\sum_2 y_{2i} = 544; \;\; \bar{y}_{2i} = 90{,}667; \;\; \sum_2 x_{2i} y_{2i} = 12868.$$

Коэффициенты регрессии для второй выборки:

$$b_2 = \frac{n_2 \sum\limits_2 x_i y_i - \sum\limits_2 x_i \sum\limits_2 y_i}{n_2 \sum\limits_2 x_i^2 - \Big(\sum\limits_2 x_i\Big)^2} =$$

$$= \frac{6 \cdot 12868 - 131 \cdot 544}{6 \cdot 3115 - 17161} = 3{,}888;$$

$$a_2 = \frac{\sum\limits_2 y_i - b_2 \sum\limits_2 x_i}{n_2} = \frac{544 - 3{,}888 \cdot 131}{6} = 5{,}78.$$

Выборочная дисперсия для x_{2i}:

$$s_{\bar{x}_2}^2 = \frac{1}{n_2 - 1} \sum_2 \Big(x_{2i} - \bar{x}_2\Big)^2 = 50{,}967;$$

$$s_{\bar{x}_2} = 7{,}139.$$

Теперь вычислим дисперсии рассеяния значений y_{1i} и y_{2i} вокруг своих линий регрессии.

$$s_1^2 = \frac{1}{n_1 - 2} \sum_1 \Big(y_{1i} - a_1 - b_1 \cdot x_{1i}\Big)^2 =$$

$$= \frac{1}{10 - 2} \sum_1 \Big(y_{1i} - 1{,}613 - 3{,}992 \cdot x_{1i}\Big)^2 = 10{,}056;$$

$$s_2^2 = \frac{1}{n_2 - 2} \sum_2 \Big(y_{2i} - a_2 - b_2 \cdot x_{2i} \Big)^2 =$$

$$= \frac{1}{6-2} \sum_2 \Big(y_{2i} - 5{,}78 - 3{,}888 \cdot x_{2i} \Big)^2 = 7{,}027.$$

Для того чтобы проверить, неразличимы ли регрессии, надо проверить выполнение трех условий (при заданной доверительной вероятности):

- $s_1^2 = s_2^2$ (равенство остаточных дисперсий);
- $a_1 = a_2$;
- $b_1 = b_2$.

Сначала проверяется равенство остаточных дисперсий с помощью критерия Фишера. Если

$$\frac{s_1^2}{s_2^2} < F_\gamma(n_1 - 2, n_2 - 2),$$

то остаточные дисперсии признаются одинаковыми. Для нашей задачи

$$\frac{s_1^2}{s_2^2} = \frac{10{,}056}{7{,}027} = 1{,}431;$$

$$F_\gamma(n_1 - 2, n_2 - 2) = F_{0{,}95}(8; 4) = 6{,}04.$$

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

$$s^2_{\bar{x}_1} = \frac{1}{n_1 - 1} \sum_1 \left(x_{1i} - \bar{x}_1 \right)^2; \quad s^2_{\bar{x}_2} = \frac{1}{n_2 - 1} \sum_2 \left(x_{2i} - \bar{x}_2 \right)^2;$$

$$\bar{x}_1 = \frac{1}{n_1} \sum_1 x_{1i}; \quad \bar{x}_2 = \frac{1}{n_2} \sum_2 x_{2i};$$

$$s^* = \sqrt{\frac{(n_1 - 2) \cdot s_1^2 + (n_2 - 2) \cdot s_2^2}{n_1 + n_2 - 4}};$$

$$s_1^2 = \frac{1}{n_1 - 2} \sum_1 \left(y_{1i} - a_1 - b_1 \cdot x_{1i} \right)^2;$$

$$s_2^2 = \frac{1}{n_2 - 2} \sum_2 \left(y_{2i} - a_2 - b_2 \cdot x_{2i} \right)^2.$$

Если

$$|t_b| \leqslant t_{k,\gamma} = T^{-1} \left(n_1 + n_2 - 4, \frac{1 + \gamma}{2} \right),$$

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

$$s^* = \sqrt{\frac{8 \cdot 10{,}056 + 4 \cdot 7{,}0271}{12}} = 3{,}008;$$

$$t_b = \frac{3{,}992 - 3{,}888}{3{,}008 \cdot \sqrt{\frac{1}{9 \cdot 88{,}988} + \frac{1}{5 \cdot 50{,}967}}} = 0{,}485.$$

По табл. 15 t-распределения Стьюдента находим

$$t_{10+6-4;0{,}95} = T^{-1}\big(12; 0{,}975\big) = 2{,}179.$$

Поскольку 0,485 < 2,179, то коэффициенты b_1 и b_2 признаются статистически равными с доверительной вероятностью $\gamma = 0{,}95$.

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:

$$\bar{b} = \frac{9 \cdot 88{,}989 \cdot 3{,}992 + 5 \cdot 50{,}967 \cdot 3{,}888}{9 \cdot 88{,}989 + 5 \cdot 50{,}967} = 3{,}967;$$

$$\tilde{b} = \frac{57{,}9 - 90{,}667}{14{,}1 - 21{,}83} = 4{,}239;$$

$$\tilde{s} = 3{,}008 \times$$

$$\times \sqrt{\frac{1}{9 \cdot 88{,}989 + 5 \cdot 50{,}967} + \frac{1}{(14{,}1 - 21{,}83)^2} \cdot \left(\frac{1}{10} + \frac{1}{6} \right)} =$$

$$= 0{,}2212.$$

Статистика t_a:

$$t_a = \frac{3{,}967 - 4{,}239}{0{,}2212} = -1{,}23.$$

Из табл. 15 $t_{12;0,95} = T^{-1}\big(12; 0{,}975\big) = 2{,}179$.

Поскольку $|-1{,}23| < 2.179$, то коэффициенты a_1 и a_2 также признаются статистически равными с доверительной вероятностью $\gamma = 0{,}95$. Таким образом, регрессии

$$y = f_1(x) = 1{,}613 + 3{,}992 \cdot x; \;\; y = f_2(x) = 5{,}78 + 3{,}888 \cdot x$$

признаются статистически идентичными.

12.6 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 в разделе 9). Это удается далеко не всегда. Кроме того, для применения методов регрессионного анализа, необходимо, чтобы функция от нормально распределенной случайной величины также оказалась нормально распределенной [11].

Рассмотрим универсальный метод построения нелинейной регрессии. Большинство нелинейных регрессионных моделей могут быть представлены как линейные по неизвестным параметрам:

$$y = y(x, \{\Theta_i\}) = \Theta_0 f_0(x) + \Theta_1 f_1(x) + \cdots + \Theta_r f_r(x),$$

где $x = \{x_1, x_2, \ldots, x_n\}; y = \{y_1, y_2, \ldots, y_n\}$ — результаты наблюдений, для которых ищется в общем случае нелинейная регрессионная связь; $\Theta_0, \Theta_1, \ldots, \Theta_r$ — неизвестные и требующие оценки параметры модели; $f_0(x), f_1(x), \ldots, f_r(x)$ — заданные функции наблюдений $\{x_i\}$.

В дальнейшем будем рассматривать разложения функции $f(x)$ только по полиномам[17] (в ряд Тейлора):

$$f_0(x) = 1;\ \ f_1(x) = x;\ \ f_2(x) = x^2;\ \ \ldots;\ \ f_r(x) = x^r.$$

При этом

$$y = \Theta_0 + \Theta_1 x + \Theta_2 x^2 + \cdots + \Theta_r x^r. \qquad (17)$$

Отметим, что если ограничиться первой степенью по x ($r = 1$) и оценивать четыре параметра $(\Theta_0, \Theta_1, \Theta_2, \Theta_3)$, то придем к задаче, аналогичной разобранной в разделе 9. Различие в том, что рассматриваемая здесь модель уже полиномиальная. В обоих случаях для нахождения неизвестных параметров модели используется МНК.

В общем случае случайная величина $y_i (i = 1, 2, \ldots, n)$ может быть представлена как

$$y_i = \Theta_0 + \Theta_1 x_i + \Theta_2 x_i^2 + \cdots + \Theta_r x_i^r + \varepsilon_i,$$

где ε_i — ошибки (невязки), случайные величины с одинаковой дисперсией, хотя распределение этих ошибок может не быть нормальным. Как и раньше, неизвестные параметры $\Theta_0, \Theta_1, \ldots, \Theta_r$ модели будем искать минимизацией по этим переменным суммы квадратов невязок[18]:

$$S = \sum \varepsilon_i^2;$$

[17] Существует много видов разложений функции $f(x)$, для которых применим нижеследующий формализм, например, разложение в ряд Фурье, когда

$$f_0(x) = 1/2;\ \ f_1(x) = \sin x;\ \ f_2(x) = \cos x;\ \ \ldots;$$

$$f_{2r-1}(x) = \sin rx;\ \ f_{2r}(x) = \cos rx.$$

[18] Здесь и далее, в сравнении с рассматриваемым ранее методом МНК применительно к нахождению решения условных уравнений, знаки «–» и «+» перед неизвестными параметрами вводятся для удобства дальнейших вычислений.

$$S = \sum \left(y_i - \Theta_0 - \Theta_1 x_i - \ldots - \Theta_r x_i^r\right)^2.$$

Обозначено как и раньше $\sum_{i=1}^{n} = \sum$.

Необходимое условие минимума задает систему нормальных уравнений:

$$\begin{cases} \dfrac{\partial S}{\partial \Theta_0} = -2 \sum \left(y_i - \Theta_0 - \Theta_1 x_i - \ldots - \Theta_r x_i^r\right) = 0 \\ \dfrac{\partial S}{\partial \Theta_1} = -2 \sum x_i \left(y_i - \Theta_0 - \Theta_1 x_i - \ldots - \Theta_r x_i^r\right) = 0 \\ \ldots \\ \dfrac{\partial S}{\partial \Theta_r} = -2 \sum x_i^r \left(y_i - \Theta_0 - \Theta_1 x_i - \ldots - \Theta_r x_i^r\right) = 0 \end{cases}.$$

Или, преобразовав:

$$\begin{cases} n\Theta_0 + \Theta_1 \sum x_i + \cdots + \Theta_r \sum x_i^r = \sum y_i \\ \Theta_0 \sum x_i + \Theta_1 \sum x_i^2 + \cdots + \Theta_r \sum x_i^{r+1} = \sum x_i y_i \\ \ldots \\ \Theta_0 \sum x_i^r + \Theta_1 \sum x_i^{r+1} + \cdots + \Theta_r \sum x_i^{2r} = \sum x_i^r y_i \end{cases}.$$

Запишем решение в матричной форме.

Введем матрицу системы (которая называется: *основная, конструкционная, структурная*[19]):

$$A = \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^r \\ 1 & x_2 & x_2^2 & \ldots & x_2^r \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_n & x_n^2 & \ldots & x_n^r \end{pmatrix}.$$

[19] Если $n = r$, то такая матрица называется матрицей Вандермонда.

Определяемые параметры и случайные ошибки представим в виде векторов:

$$\Theta^T = \big(\Theta_0, \Theta_1, \Theta_2, \ldots, \Theta_r\big);$$

$$\varepsilon^T = \big(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n\big).$$

В матричном обозначении:

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^T \cdot \varepsilon =$$

$$= \big(y - A\Theta\big)^T \cdot \big(y - A\Theta\big) = y^T y - 2\Theta^T A^T y + \Theta^T A^T A\Theta.$$

По методу МНК нужно дифференцировать последнее равенство по всем параметрам Θ_i и приравнять результат к нулю:

$$-2A^T y + 2A^T A\Theta = 0,$$

что можно переписать в виде

$$\big(A^T A\big)\Theta = A^T y.$$

Решение последнего уравнения:

$$\tilde{\Theta} = \big(A^T A\big)^{-1} A^T y. \tag{18}$$

12.6.1 Ортогональные полиномы и преимущества их использования

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,

$$\bar{x} = \frac{1}{n} \sum x_i.$$

Здесь и далее в этом разделе обозначено $\sum_{i=1}^{n} = \sum$.

Рассмотрим значительные преимущества такой замены переменных. Аналогично предыдущим выводам в матричной форме уравнения МНК имеют вид

$$y = B\Phi + \varepsilon,$$

где новая конструкционная матрица B:

$$B = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix}.$$

Поскольку полученная модель — линейная (если положить $f_0(x) = 1; f_1(x) = x - \bar{x}$), то ее решение имеет вид, полностью аналогичный (18):

$$\tilde{\Phi} = \left(B^T B\right)^{-1} B^T y, \tag{19}$$

где

$$B^T B = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 - \bar{x} & x_2 - \bar{x} & \dots & x_n - \bar{x} \end{pmatrix} \cdot \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix} =$$

$$= \begin{pmatrix} n & \sum x_i - n\bar{x} \\ \sum x_i - n\bar{x} & \sum \left(x_i - \bar{x}\right)^2 \end{pmatrix} = \begin{pmatrix} n & 0 \\ 0 & \sum \left(x_i - \bar{x}\right)^2 \end{pmatrix}.$$

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 (17):

$$y = \Theta_0 + \Theta_1 x + \Theta_2 x^2 + \cdots + \Theta_r x^r.$$

Такая модель приводит к матрице

$$A^TA = \begin{pmatrix} n & \sum x_i & \sum x_i^2 & \dots & \sum x_i^r \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \dots & \sum x_i^{r+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_i^r & \sum x_i^{r+1} & \sum x_i^{r+2} & \dots & \sum x_i^{2r} \end{pmatrix}, \quad (20)$$

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. Пусть множество $n\{x_i\}$ равномерно расширяется при росте $n \to \infty$. Тогда

$$\sum x_i^r \to n \int_0^1 x^r dx = \frac{n}{r+1}.$$

Следовательно,

$$A^T A \approx$$

$$\approx n \cdot \begin{pmatrix} 1 & 1/2 & 1/3 & \dots & 1/(r+1) \\ 1/2 & 1/3 & 1/4 & \dots & 1/(r+2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1/(r+1) & 1/(r+2) & 1/(r+3) & \dots & 1/(2r+1) \end{pmatrix} =$$

$$= nH_{r+1},$$

где H_p — гильбертова матрица ранга p, которая плохо обусловлена при больших p [4].

Для того, чтобы избежать таких неустойчивостей решения, вместо модели (17) вводится эквивалентная ей модель, записанная в виде ортогональных полиномов:

$$\xi = \xi(x, \Phi_0, \Phi_1, \dots, \Phi_r) =$$

$$= \Phi_0 + \Phi_1 G_1(x) + \Phi_2 G_2(x) + \dots + \Phi_r G_r(x),$$

где полиномы

$$G_j(x) = k_j^{(0)} + k_j^{(1)} x + \dots + k_j^{(j-1)} x^{j-1} + x^j.$$

Для рассмотренного линейного случая

$$G_1(x) = 1;\ G_2(x) = x - \bar{x}.$$

В общем случае коэффициенты $k_j^{(0)}, k_j^{(1)}, \dots, k_j^{(j-1)}$ определяются из системы (метод ортогонализации Грама–Шмидта):

$$\begin{cases} \sum G_0(x_i) \cdot G_j(x_i) = 0 \\ \\ \sum G_1(x_i) \cdot G_j(x_i) = 0 \\ \\ \dots \\ \\ \sum G_{j-1}(x_i) \cdot G_j(x_i) = 0 \end{cases};$$

$$G_0(x) = 1.$$

Свойство ортогональности означает, что

$$\int_0^1 G_j(x)G_m(x)dx = 0 \ \ (j \neq m).$$

Последнее выражение означает, что все недиагональные элементы матрицы $B^T B$ обращаются в ноль:

$$B^T B =$$

$$= \begin{pmatrix} \sum G_0^2 & \sum G_0 G_1 & ... & \sum G_0 G_r \\ \sum G_1 G_0 & \sum G_1^2 & ... & \sum G_1 G_r \\ \vdots & \vdots & \ddots & \vdots \\ \sum G_r G_0 & \sum G_r G_1 & ... & \sum G_r^2 \end{pmatrix} =$$

$$= \begin{pmatrix} \sum G_0^2 & 0 & ... & 0 \\ 0 & \sum G_1^2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \sum G_r^2 \end{pmatrix}_{(r+1)(r+1)},$$

где все $G_0, G_1, ..., G_r$ — функции от x_i.

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:

$$\sum G_{r+1}^2(x_i).$$

12.6.2 Ортогональные нормированные (ортонормальные) полиномы и преимущества их использования

Следующий шаг для упрощения аппроксимации — использование не просто ортогональных, но *нормированных ортогональных полиномов*. Для краткости ортогональные нормированные полиномы называют *ортонормированными* или *ортонормальными*.

Вместо метода Грама–Шмидта используется метод Форсайта, который заключается в установлении простого рекуррентного соотношения между ортонормальными полиномами вида

$$Q_j(x) = \frac{G_j(x)}{\sqrt{\sum_{i=1}^{n} G_j^2(x_i)}};$$

$$Q_0(x) = \frac{1}{\sqrt{n}}; \quad \sum_{i=1}^{n} Q_j^2(x_i) = 1.$$

Пусть разложение функции $f(x)$ по ортонормальным полиномам имеет вид:

$$\tilde{f}(x) = \omega_0 Q_0(x) + \omega_1 Q_1(x) + \cdots + \omega_r Q_r(x). \qquad (21)$$

Для $Q_i(x)$ выполняется рекуррентное соотношение:

$$\lambda Q_j(x) = xQ_{j-1}(x) - \alpha Q_{j-1}(x) - \beta Q_{j-2}(x),$$

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 метод МНК к выражению (21), получаем формулу, аналогичную выражению (19):

$$\tilde{\omega} = \left(B^T B\right)^{-1} B^T y = B^T y,$$

поскольку для ортонормальных полиномов $A^T A = I$. Матрица B:

$$B = \begin{pmatrix} Q_0(x_1) & Q_1(x_1) & \dots & Q_r(x_1) \\ Q_0(x_2) & Q_1(x_2) & \dots & Q_r(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ Q_0(x_n) & Q_1(x_n) & \dots & Q_r(x_n) \end{pmatrix}.$$

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 для (21):

$$\tilde{f}(x) = \omega_0 Q_0(x) + \omega_1 Q_1(x) + \dots + \omega_r Q_r(x) =$$

$$= \tilde{\omega}_0 Q_0(x) + \tilde{\omega}_1 Q_1(x) + \cdots + \tilde{\omega}_r Q_r(x).$$

Остаточная сумма квадратов (невязка или остаточная погрешность) при аппроксимации ортонормальными полиномами степени r:

$$\varepsilon_r^2 = \sum_{i=1}^{n} y_i^2 - \sum_{j=0}^{r} \left(\sum_{i=1}^{n} y_i Q_j(x_i) \right)^2,$$

а остаточная дисперсия:

$$\sigma_\varepsilon^2 = \frac{\varepsilon_r^2}{n-r-1}.$$

12.6.3 Правила вычисления ортонормальных полиномов Чебышёва на дискретном наборе точек

Коэффициенты разложения $\tilde{\omega}_i$ по 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. В качестве примера рассмотрим *полиномы Чебышёва* [3]. Особенностью использования полиномов Чебышёва является требование того, чтобы все точки $x = \{x_i\}$ были равноотстоящими (эквидистантными).

Сначала рассмотрим систему ортогональных, но ненормированных полиномов Чебышёва.

Пусть дана система $n+1$ равноотстоящих точек

$$x = \{x_i\} \ \ (i = 0, 1, 2, \ldots, n).$$

С помощью линейного преобразования $t = (x - x_0)/h$ переведем эти точки в $t = 0, 1, 2, \ldots, n$ соответственно.

Полиномы $P_{0,n}(t), P_{1,n}(t), \ldots, P_{m,n}(t) (m \leqslant n)$ степеней $0, \ldots, m$, ортогональные на множестве точек $\{0, 1, 2, \ldots, n\}$ и отличные от нуля на этом множестве, называются *ортогональными полиномами Чебышёва*. Первый индекс в $P_{k,n}(t)$, k — степень полинома, а второй индекс n — число точек, уменьшенное на единицу.

Полиномы Чебышёва задаются формулой

$$P_{k,n}(t) = \sum_{s=0}^{k} (-1)^s C_k^s C_{k+s}^s \cdot \frac{t^{[s]}}{n^{[s]}} \ (k = 0, 1, 2, \ldots, m), \quad (22)$$

где

$$C_n^k = \frac{n!}{k!(n-k)!}.$$

Здесь введено понятие *обобщенной степени*:

$$t^{[s]} = t(t-1)\ldots(t-s+1) \equiv \frac{t!}{(t-s)!};$$

$$n^{[s]} = n(n-1)\ldots(n-s+1) \equiv \frac{n!}{(n-s)!}.$$

Укажем, как может быть получен удобный для решения практических задач вид (22) [3]. Полином Чебышёва есть, по определению, такой многочлен

$$P_{k,n}(t) = 1 + b_1 t + b_2 t^{[2]} + b_3 t^{[3]} + \cdots + b_k t^{[k]}, \quad (23)$$

что

$$\sum_{t=0}^{n} (t+s)^{[s]} P_{k,n}(t) = 0 \ \ (s = 0, 1, 2, \ldots, k-1).$$

Умножая равенство (23) на $(t+s)^{[s]}$, получим

$$(t+s)^{[s]} P_{k,n}(t) = (t+s)^{[s]} + b_1 (t+s)^{[s+1]} + \ldots b_k (t+s)^{[s+k]}.$$

Суммируем последнее выражение по t от 0 до n:

$$0 = \sum_{t=0}^{n} (t+s)^{[s]} P_{k,n}(t) =$$

$$= \sum_{t=0}^{n} \left((t+s)^{[s]} + b_1 (t+s)^{[s+1]} + \dots b_k (t+s)^{[s+k]} \right) =$$

$$= \sum_{t=0}^{n} (t+s)^{[s]} + \sum_{t=0}^{n} b_1 (t+s)^{[s+1]} + \cdots \sum_{t=0}^{n} b_k (t+s)^{[s+k]} =$$

$$= \frac{(n+s+1)^{[s+1]}}{s+1} + b_1 \frac{(n+s+1)^{[s+2]}}{s+2} + \cdots +$$

$$+ b_k \frac{(n+s+1)^{[s+k+1]}}{s+k+1}.$$

Последнее равенство можно доказать по индукции относительно n и для любого s.

Заметим, что

$$\sum_{t=0}^{n} (t+s)^{[s]} = \sum_{t=0}^{n} \frac{(t+s)!}{t!};$$

$$\frac{(n+s+1)^{[s+1]}}{s+1} = \frac{(n+s+1)!}{n!(s+1)}.$$

Для $n = 0$ и любого s равенство верно:

$$\sum_{t=0}^{n} \frac{(t+s)!}{t!} = \frac{(n+s+1)!}{n!(s+1)}.$$

Пусть последнее равенство верно для некоторого n. Докажем, что оно верно и для $n+1$:

$$\sum_{t=0}^{n+1} \frac{(t+s)!}{t!} = \frac{(n+s+2)!}{(n+1)!(s+1)}.$$

Действительно,

$$\sum_{t=0}^{n+1} \frac{(t+s)!}{t!} =$$

$$s! + \frac{(s+1)!}{1!} + \frac{(s+2)!}{2!} + \cdots + \frac{(s+n)!}{n!} + \frac{(s+n+1)!}{(n+1)!} =$$

$$= \frac{(n+s+1)!}{n!(s+1)} + \frac{(s+n+1)!}{(n+1)!} =$$

$$= \frac{(n+s+1)!}{(n+1)!} \cdot \left(\frac{n+1}{s+1} + 1\right) =$$

$$= \frac{(n+s+1)!}{(n+1)!} \cdot \frac{n+s+2}{s+1} = \frac{(n+s+2)!}{(n+1)!(s+1)},$$

что и требовалось доказать[20].

Аналогичным образом по индукции можно доказать:

$$\sum_{t=0}^{n} b_1 (t+s)^{[s+1]} = b_1 \frac{(n+s+1)^{[s+2]}}{s+2};$$

$$\ldots;$$

$$\sum_{t=0}^{n} b_k (t+s)^{[s+k]} = b_k \frac{(n+s+1)^{[s+k+1]}}{s+k+1}.$$

Далее разделим на выражение $(n+s+1)^{[s+1]}$ следующее равенство:

$$\frac{(n+s+1)^{[s+1]}}{s+1} +$$

[20] В частном случае при $s = 2$ получаем известное выражение суммы квадратов $(n+1)$ целых чисел:

$$\sum_{t=0}^{n} t^2 = \frac{n(n+1)(2n+1)}{6}.$$

$$+b_1\frac{(n+s+1)^{[s+2]}}{s+2}+\cdots+b_k\frac{(n+s+1)^{[s+k+1]}}{s+k+1}=0;$$

После преобразования получаем:

$$\frac{1}{s+1}+\frac{b_1 n}{s+2}+\frac{b_2 n^{[2]}}{s+3}+\cdots+\frac{b_k n^{[k]}}{s+k+1}=0. \qquad (24)$$

Выражение (24) представляет собой группу уравнений относительно коэффициентов b_j. Их решения:

$$b_j=\big(-1\big)^j C_j^k C_j^{k+j}\cdot\frac{1}{n^{[j]}}.$$

Получить эти решения можно следующим образом. Нужно привести (24) к общему знаменателю, в результате чего в числителе получится многочлен, который, исходя из (24), должен обращаться в 0 при $s=0,1,\ldots,k-1$. Следовательно, этот многочлен имеет вид: $\mathrm{const}\cdot s^{[k]}$. Далее, последовательно полагая s равным различным отрицательным числам, находим эту постоянную и все b_j, подставляя которые в (23), получаем (22).

Несколько первых ортогональных полиномов Чебышёва:

$$P_{0,n}(t)=1;$$

$$P_{1,n}(t)=1-2\cdot\frac{t}{n}\ \ (n\geqslant 1);$$

$$P_{2,n}(t)=1-6\cdot\frac{t}{n}+6\cdot\frac{t(t-1)}{n(n-1)}\ \ (n\geqslant 2);$$

$$P_{3,n}(t)=1-12\cdot\frac{t}{n}+30\cdot\frac{t(t-1)}{n(n-1)}-20\cdot\frac{t(t-1)(t-2)}{n(n-1)(n-2)}\ \ (n\geqslant 3);$$

$$P_{4,n}(t)=1-20\cdot\frac{t}{n}+90\cdot\frac{t(t-1)}{n(n-1)}-140\cdot\frac{t(t-1)(t-2)}{n(n-1)(n-2)}+$$
$$+70\cdot\frac{t(t-1)(t-2)(t-3)}{n(n-1)(n-2)(n-3)}\ \ (n\geqslant 4).$$

Возвращаясь к прежней переменной x, получим систему полиномов, ортогональных на дискретном множестве $x = \{x_i\}$:

$$P_{k,n}\left(\frac{x - x_0}{h}\right) \ (k = 0, 1, \ldots, m; \ \ m \leqslant n).$$

Система полиномов $\{P_{k,n}(t)\}$ не является нормированной. Построим соответствующую нормированную систему и получим ортонормальные полиномы Чебышёва. Определим *норму* для $\{P_{k,n}(t)\}$ следующим образом [3]:

$$\left\|P_{k,n}(t)\right\|^2 = \sum_{i=0}^{n} P_{k,n}^2(i) = \frac{\left(n + k + 1\right)^{[k+1]}}{\left(2k + 1\right) \cdot n^{[k]}}.$$

Последнее выражение может быть выведено из условия ортогональности полиномов Чебышёва:

$$\sum_{i=0}^{n} P_{m,n}(i) P_{k,n}(i) = 0 \ \ (m \neq k).$$

Разделив многочлены $P_{k,n}(t)$ на их нормы, получим *ортонормальную систему полиномов Чебышёва*:

$$\tilde{P}_{k,n}(t) = \frac{P_{k,n}(t)}{\left\|P_{k,n}(t)\right\|} \ \ (k = 0, 1, 2, 3 \ldots, m; \ \ m \leqslant n).$$

ПРИМЕР. Пусть задана система точек:

$$x_0 = 1/2; x_1 = 1; x_2 = 3/2; x_3 = 2; x_4 = 5/2; x_5 = 3.$$

Построим систему полиномов Чебышёва до 3-й степени включительно, ортонормальную 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:

$$t = \frac{x - x_0}{h} = \frac{x - 1/2}{1/2} = 2 \cdot \big(x - 1/2\big),$$

при которой все x_i переходят в целочисленные значения $t = 0, 1, 2, 3, 4, 5$. В общей формуле

$$P_{k,n}(t) = \sum_{s=0}^{n} \big(-1\big)^s C_k^s C_{k+s}^s \cdot \frac{t^{[s]}}{n^{[s]}}$$

примем $n = 5$, тогда

$$P_{k,5}(t) = \sum_{s=0}^{5} \big(-1\big)^s C_k^s C_{k+s}^s \cdot \frac{t^{[s]}}{5^{[s]}},$$

где $k = 0, 1, 2, 3$, поскольку по условию ищется аппроксимирующий полином степени $m = 3$. Проведем вычисления для всех указанных k:

$$P_{0;5}(t) = \sum_{s=0}^{5} \big(-1\big)^s C_0^s C_s^s \cdot \frac{t^{[s]}}{5^{[s]}} = 1;$$

$$P_{1;5}(t) = \sum_{s=0}^{5} \big(-1\big)^s C_1^s C_{1+s}^s \cdot \frac{t^{[s]}}{5^{[s]}} = 1 - 0{,}4 \cdot t;$$

$$P_{2;5}(t) = \sum_{s=0}^{5} \big(-1\big)^s C_2^s C_{2+s}^s \cdot \frac{t^{[s]}}{5^{[s]}} = 1 - 1{,}2 \cdot t + 0{,}3 \cdot t(t-1);$$

$$P_{3;5}(t) = \sum_{s=0}^{5} \big(-1\big)^s C_3^s C_{3+s}^s \cdot \frac{t^{[s]}}{5^{[s]}} = 1 - 2{,}4 \cdot t + 1{,}5 \cdot t(t-1) -$$

$$- 0{,}333 \cdot t(t-1)(t-2).$$

Нормы функций $P_{k,5}(t)$, где $k = 0, 1, 2, 3$, вычисляем по формуле

$$\left\|P_{k,n}(t)\right\|^2 = \frac{\left(n+k+1\right)^{[k+1]}}{\left(2k+1\right)\cdot n^{[k]}}:$$

$$\left\|P_{0;5}(t)\right\| = \sqrt{6};$$

$$\left\|P_{1;5}(t)\right\| = \sqrt{\frac{7\cdot 6}{3\cdot 5}} = \sqrt{\frac{14}{5}};$$

$$\left\|P_{2;5}(t)\right\| = \sqrt{\frac{8\cdot 7\cdot 6}{5\cdot 5\cdot 4}} = \frac{2}{5}\sqrt{21};$$

$$\left\|P_{3;5}(t)\right\| = \sqrt{\frac{9\cdot 8\cdot 7\cdot 6}{7\cdot 5\cdot 4\cdot 3}} = \frac{6}{\sqrt{5}}.$$

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 Чебышёва:

$$\tilde{P}_{0;5}(x) = \frac{1}{\sqrt{6}} = 0{,}408;$$

$$\tilde{P}_{1;5}(x) = \sqrt{\frac{5}{14}}\left(1-0{,}8\cdot(x-1/2)\right) = 0{,}837-0{,}478\cdot x;$$

$$\tilde{P}_{2;5}(x) = \frac{5}{2\sqrt{21}}\left(1-2{,}4\cdot(x-1/2)+0{,}6\cdot(x-1/2)(x-3/2)\right) =$$

$$= 1{,}446-1{,}964\cdot x+0{,}327\cdot x^2;$$

$$\tilde{P}_{3;5}(x) =$$

$$= \frac{\sqrt{5}}{6}\left(1-4{,}8\cdot(x-1/2)+3\cdot(x-1/2)(x-3/2)-\right.$$

$$-0{,}666 \cdot (x-1/2)(x-3/2)(x-5/2)\Big) =$$

$$= 2{,}571 - 5{,}452 \cdot x + 2{,}235 \cdot x^2 - 0{,}248 \cdot x^3.$$

12.6.4 Нахождение уравнения регрессии с помощью ортонормальных полиномов Чебышёва и определение порядка нелинейности

Если функция $y = f(x)$ задана на множестве узлов $x = \{x_0, x_1, \ldots, x_n\}$ с шагом h, то наилучший (по МНК) аппроксимирующий полином ищется в виде

$$\tilde{f}(x) = \sum_{k=0}^{m} \omega_k \cdot P_{k,n}\left(\frac{x-x_0}{h}\right),$$

где коэффициенты ω_k называются *коэффициентами Фурье* функции $f(x)$ относительно системы ортогональных полиномов Чебышёва $P_{k,n}((x-x_0)/h)$ $(k = 0, 1, 2, \ldots, m)$:

$$\omega_k = \frac{\sum_{i=0}^{n} y_i \cdot P_{k,n}(i)}{\left\|P_{k,n}(t)\right\|^2}.$$

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 $B^T y$:

$$\tilde{\omega}_k = \sum_{i=0}^{n} y_i \cdot \tilde{P}_{k,n}(i).$$

Остается вопрос о статистическом критерии выбора степени аппроксимирующего полинома m.

Остаточная дисперсия при аппроксимации ортонормальными полиномами Чебышёва степени m:

$$\varepsilon_m^2 = \frac{\sum\limits_{i=0}^{n} y_i^2 - \sum\limits_{j=0}^{m}\left(\sum\limits_{i=0}^{n} y_i \tilde{P}_{j,n}(x_i)\right)^2}{(n+1)-m-1}.$$

Так, если

$$\frac{\varepsilon_{m+1}^2}{\varepsilon_m^2} > 1,$$

то в качестве регрессии принимается полином степени m. Значимость отличия остаточных дисперсий на каждом шаге увеличения степени полинома дается критерием Фишера $F_\gamma(n-m, n-m-1)$. Так, если

$$\frac{\varepsilon_1^2}{\varepsilon_2^2} > F_\gamma\big((n+1)-2, (n+1)-2-1\big),$$

то полином второй степени ($m = 2$) предпочтительнее полинома первой степени (квадратичная регрессия предпочтительнее линейной). Если

$$\frac{\varepsilon_2^2}{\varepsilon_3^2} > F_\gamma\big((n+1)-3, (n+1)-3-1\big),$$

то полином третьей степени ($m = 3$) предпочтительнее полинома второй степени.

ПРИМЕР. Пусть в результате наблюдений получены пары $\{x_i, y_i\}$ (см. табл. 28).

Поскольку $x_{i+1} - x_i = h = 2{,}1$, введем замену переменной:

$$t = \frac{x - 1{,}1}{2{,}1}.$$

Таблица 28

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 $\{x_i\}(i = 1, 2, \ldots, 10)$ систему ортогональных полиномов Чебышёва ($n = 9$, 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:

$$\left\| P_{0;9}(t) \right\| = \sqrt{10};$$

$$\left\| P_{1;9}(t) \right\| = \sqrt{\frac{11 \cdot 10}{3 \cdot 9}} = \sqrt{\frac{110}{27}};$$

$$\left\| P_{2;9}(t) \right\| = \sqrt{\frac{12 \cdot 11 \cdot 10}{5 \cdot 9 \cdot 8}} = \sqrt{\frac{11}{3}};$$

$$\left\|P_{3;9}(t)\right\| = \sqrt{\frac{13\cdot 12\cdot 11\cdot 10}{7\cdot 9\cdot 8\cdot 7}} = \sqrt{\frac{715}{147}};$$

$$\left\|P_{4;9}(t)\right\| = \sqrt{\frac{14\cdot 13\cdot 12\cdot 11\cdot 10}{9\cdot 9\cdot 8\cdot 7\cdot 6}} = \sqrt{\frac{715}{81}}.$$

Окончательно ортонормальная система полиномов (до 4-го порядка) имеет вид:

$$\tilde{P}_{0;9}(t) = 0{,}3162;\ \ \tilde{P}_{1;9}(t) = 0{,}4954 - 0{,}1101\cdot t;$$

$$\tilde{P}_{2;9}(t) = 0{,}5222 - 0{,}3917\cdot t + 0{,}0435\cdot t^2;$$

$$\tilde{P}_{3;9}(t) = 0{,}4534 - 0{,}8295\cdot t + 0{,}2429\cdot t^2 - 0{,}0180\cdot t^3;$$

$$\tilde{P}_{4;9}(t) = 0{,}3366 - 1{,}4024\cdot t + 0{,}7869\cdot t^2 - 0{,}1402\cdot t^3 + 0{,}0078\cdot t^4.$$

Отметим, что для соблюдения необходимой в конкретной задаче точности следует выполнять действия с иррациональными выражениями, округляя до нужного знака только окончательный результат.

Вычислим остаточные дисперсии для $m = 1, 2, 3, 4$ по формуле:

$$\varepsilon_m^2 = \frac{\displaystyle\sum_{i=0}^{n} y_i^2 - \sum_{j=0}^{m}\left(\sum_{i=0}^{n} y_i \tilde{P}_{j,n}(x_i)\right)^2}{(n+1) - m - 1}:$$

$$\varepsilon_1^2 = 7511{,}5068;\ \ \varepsilon_2^2 = 378{,}7319;$$

$$\varepsilon_3^2 = 0{,}0007;\ \ \varepsilon_4^2 = 0{,}0008.$$

Как видно из вычислений, с увеличением степени аппроксимирующего полинома остаточная дисперсия в целом падает, т.е. аппроксимация становится все точнее. В данном примере отметим значительный скачок остаточной дисперсии при аппроксимации полиномом второй и третьей степеней, который означает, что аппроксимация

данных параболой существенно хуже, чем аппроксимация полиномом третьей степени.

Сравним, значимы ли различия остаточных дисперсий:

$$\frac{\varepsilon_1^2}{\varepsilon_2^2} = 19{,}833 > F_{0,95}(8;7) = 3{,}73;$$

$$\frac{\varepsilon_2^2}{\varepsilon_3^2} \approx 10^5 > F_{0,95}(7;6) = 4{,}21;$$

но

$$\frac{\varepsilon_3^2}{\varepsilon_4^2} = 0{,}845 < F_{0,95}(6;5) = 4{,}95,$$

что означает, что различие остаточных дисперсий при аппроксимации полиномами 3-го и 4-го порядков незначимо. Таким образом, наблюдательные данные аппроксимируются нелинейной регрессионной моделью в виде полинома 3-го порядка с доверительной вероятностью $\gamma = 0{,}95$.

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:

$$\tilde{f}(x) = 0{,}8049 - 0{,}3067 \cdot x + 0{,}8010 \cdot x^2 - 0{,}1000 \cdot x^3.$$

13 Исследование вида распределения

Пусть $\{x_i\}$ $(i = 1, 2, \ldots n)$ 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 данных на соответствие функции $F(x)$ общего вида, а далее, в примерах, конкретизируем вид функции $F(x)$.

13.1 Критерий χ^2 (хи-квадрат)

Критерий χ^2 (или *критерий Пирсона*) позволяет количественно оценить отклонение наблюдательных и экспериментальных данных от теоретического распределения известной структуры, но с неизвестными параметрами. Мерой отклонения является величина, которая используется для построения доверительной области для неизвестной плотности распределения. Производится замена неизвестных истинных значений вероятностей попадания в интервалы вероятностями, вычисленными по теоретическому распределению.

Алгоритм проверки соответствия случайной выборки заданной функции распределения строится с помощью критерия χ^2 [10].

По выборке наблюдений находят оценки неизвестных параметров предполагаемого закона распределения случайной величины X. Далее область возможных значений случайной величины X разбивается на r подмножеств

$\Delta_1, \Delta_2, \ldots, \Delta_r$, например, r интервалов в случае, когда X есть 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 $(k = 1, 2, \ldots, r)$. 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:

$$p_k = P(X \in \Delta_k) \ \ (k = 1, 2, \ldots, r).$$

Очевидно, $\sum_{k=1}^{r} p_k = 1$. Полученные результаты можно представить в табл. 29.

Выборочное значение статистики критерия χ^2 есть:

$$\tilde{\chi}^2 = \sum_{k=1}^{r} \frac{\left(n_k - n \cdot p_k\right)^2}{n \cdot p_k}.$$

Пусть задана доверительная вероятность γ. Тогда предложенный закон распределения соответствует генеральной совокупности исследуемой выборки, если выполняется неравенство

$$\tilde{\chi}^2 < \chi^2_\gamma(r - l - 1),$$

где $\chi^2_\gamma(r - l - 1)$ — γ-квантиль χ^2-распределения (см. табл. 16) с $r - l - 1$ степенями свободы; 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 29

Оформление элементов выборки для проверки на соответствие заданной функции распределения.

Интервал	Δ_1	Δ_2	...	Δ_r	Контрольная сумма элементов
Число наблюдений	n_1	n_2	...	n_r	$\sum_{i=1}^{r} n_i = n$
Ожидаемое число наблюдений	np_1	np_2	...	np_r	$\sum_{i=1}^{r} n \cdot p_i = n$

Отметим важный момент. Критерий χ^2 использует тот факт, что случайная величина $(n_k - n \cdot p_k)/\sqrt{n \cdot p_k}$ $(k = 1, 2, \ldots, r)$, имеет распределение, близкое к стандартному нормальному. Чтобы это утверждение было достаточно точным, рекомендуется, чтобы для всех интервалов выполнялось условие

$$n \cdot p_k \geqslant 5.$$

Если для некоторых интервалов это условие не выполняется, то их следует объединить с соседними до выполнения этого условия.

ПРИМЕР. Рассмотрим задачу проверки выборки на соответствие распределению Пуассона. В первых двух столбцах табл. 30 приведены данные об отказах аппаратуры за 10^4 часов работы. Общее число обследованных

Таблица 30

Оформление элементов выборки для проверки на соответствие функции распределения Пуассона.

Число отказов	Количество случаев, в которых наблюдалось k отказов n_k	$p_k = \frac{0{,}6^k}{k!} \cdot \exp\{-0{,}6\}$	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
$\geqslant 6$	0	0,00004	0
Сумма	757		

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. Ставится задача проверить, распределено ли число отказов по закону Пуассона, приняв доверительную вероятность $\gamma = 0{,}99$. Вероятность для дискретного распределения Пуассона есть

$$p_k = P(X = k) = \frac{\lambda^k}{k!} \cdot \exp\{-\lambda\} \quad (k = 0, 1, 2, \dots).$$

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 (см. 3-й и 4-й столбцы табл. 30).

Таблица 31

Оформление элементов выборки для проверки на соответствие функции распределения Пуассона после объединения малочисленных 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
$\geqslant 3$	23	17	2,118
			$\tilde{\chi}^2 = 3{,}316$

Для $k = 4, 5$ и 6 значения $n \cdot p_k < 5$, поэтому объединим эти строки со строкой для $k = 3$. В результате получим значения, указанные в табл. 31.

Так как по выборке оценивается один параметр λ, то $l = 1$ и число степеней свободы равно $4 - 1 - 1 = 2$.

По табл. 16 распределения хи-квадрат находим $\chi^2_{0,99}(2) = 9{,}21$.

Вычисленная статистика $\tilde{\chi}^2 = 3{,}316 < 9{,}21$, следовательно, принимается предположение о распределении Пуассона.

ПРИМЕР. Рассмотрим задачу проверки выборки на соответствие нормальному распределению. Дана выборка из 55 наблюдений (см. табл. 32).

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 и все вычисленные величины

Таблица 32

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	

сведены в табл. 33.

В 4-м столбце табл. 33 приводятся вероятности, вычисленные по формуле:

$$p_k = P(X \in \Delta_k) = \Phi\left(\frac{\beta_k - \bar{x}}{s}\right) - \Phi\left(\frac{\alpha_k - \bar{x}}{s}\right) \quad (k = 1, \ldots, 7).$$

Здесь α_k и β_k — нижняя и верхняя границы интервалов соответственно, а значения функции стандартного нормального распределения берутся из статистической таблицы (см. табл. 13).

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 доверительной вероятностью 0,90, по статистической таблице распределения хи-квадрат (см. табл. 16) находим $\chi^2_{0,90}(2) = 4{,}61$. Округленное выборочное значение статистики критерия есть $\tilde{\chi}^2 \approx 0{,}77 < 4{,}61$, следовательно, предположение о нормальном распределении верно.

Таблица 33

Выборка, проверяемая на соответствие нормальному закону распределения $n = 55$ (во 2-м и 3-м столбце приведены результаты группировки по k интервалам, в 4-м столбце приведены вычисленные вероятности (см. текст). В 5-м столбце приводятся ожидаемые частоты, а в 6-м — значения ожидаемых частот после объединения первых двух и последних двух интервалов).

k	Δ_k	Набл. част. n_k	Вероятность попадания в интерв. Δ_k (p_k)	Ожид. част. $n \cdot p_k$	$n \cdot p_k$	$n_k - n \cdot p_k$	$\frac{(n_k - n \cdot p_k)^2}{n \cdot p_k}$
1	$(-\infty, 12)$	2	0,022958	1,262675			
2	$[12; 14)$	4	0,070648	3,885617	5,148293	0,851707	0,140902
3	$[14; 16)$	8	0,166967	9,183185	9,183185	$-1,183185$	0,152445
4	$[16; 18)$	12	0,253672	13,951939	13,951939	$-1,951939$	0,273085
5	$[18; 20)$	15	0,247835	13,630922	13,630922	1,369078	0,137509
6	$[20; 22)$	11	0,155702	8,563618	13,085662	0,914338	0,063888
7	$[22; +\infty)$	3	0,082219	4,522043			
	Сумма	55	1,0001	55,0	55,0		0,767829

13.2 Критерий Колмогорова

Критерий χ^2 — один из лучших, применяемых на практике. Однако помимо него существуют еще несколько критериев для проверки эмпирических распределений на соответствие заданному теоретическому закону распределения.

В качестве примера приведем *критерий Колмогорова* (или *Колмогорова–Смирнова*).

Мерой сравнения эмпирического и теоретического распределений выбирается расстояние

$$D_n = \max_{|x|<\infty} \Big| F_n^*(x) - F(x) \Big|,$$

где $F_n^*(x) = F_n^*(x, \vartheta)$ — эмпирическая функция распределения, а $F(x) = F(x, \vartheta)$ — теоретическая функция распределения. Параметры ϑ — это параметры распределения, неизвестные в общем случае и требующие оценок по эмпирическим данным.

А.Н. Колмогоровым было доказано, что закон распределения величины

$$\Lambda = D_n \sqrt{n} = \sqrt{n} \cdot \max_{|x|<\infty} \Big| F_n^*(x) - F(x) \Big| \tag{25}$$

при $n \to \infty$ определяется функцией

$$F(\lambda) = \sum_{k=-\infty}^{+\infty} (-1)^k \exp\left\{-2k^2\lambda^2\right\}. \tag{26}$$

Если величина Λ, вычисленная по данным выборки по формуле (25), окажется меньше, чем статистика λ_γ, которая есть табличная величина, определяемая своей функцией распределения $F(\lambda)$ (26), см. табл. 34, то эмпирическое распределение признается соответствующим теоретическому.

Таблица 34

Квантили λ_γ распределения Колмогорова $F(\lambda)$.

$F(\lambda_\gamma) = \gamma$	0,80	0,90	0,95	0,98	0,99	0,999
λ_γ	1,073	1,224	1,358	1,520	1,627	1,950

Для эффективного применения критерия Колмогорова необходимо, чтобы все параметры ϑ теоретической функции распределения $F(x, \vartheta)$ были известны заранее.

14 Непараметрические критерии

14.1 Понятие ранговых критериев

Для сравнения двух выборок, законы распределения которых неизвестны или 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 (количественный анализ того, как сильно смещены средние и искажены распределения двух выборок 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 [6] (далее эффективность везде указывается относительно t-критерия). Ранговые критерии основываются на использовании рангов, приписываемых значениям случайных величин в общей упорядоченной по возрастанию выборке. Другими словами, для двух выборок размером m и n в упорядоченном ряду чисел $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_{n+m}$ значению x_i приписывается ранг $R_i = i$. При этом одинаковым величинам присваивается усредненный ранг. Таким образом, анализируются не сами значения выборочных элементов, а их ранги, как они перемежаются для выборок в едином ряду.

14.2 Постановка задачи поиска космических струн с помощью ранговых критериев

На примере обработки реальных данных рассмотрим несколько ранговых критериев: быстрый ранговый критерий, критерий ван дер Вардена и критерий Манна–Уитни–Вилкоксона, частным случаем которого является аппроксимация Имана.

Реальные наблюдательные данные были получены в ходе поисков космической струны — гипотетического одномерного объекта космологических масштабов, существование которого следует из теорий эволюции ранней Вселенной. Согласно теории, космические струны обладают рядом хорошо описанных астрофизических свойств, благодаря которым они могут быть обнаружены. Поскольку такие объекты движутся с релятивистскими скоростями, то они могут проявлять себя посредством эффекта Доплера в реликтовом микроволновом излучении. Кроме того, влияя на глобальную структуру пространства-времени, космические струны должны порождать характерные цепочки гравитационно-линзовых событий, статистическое распределение которых исследуется с помощью непараметрических критериев. Известно, что обычные события гравитационного линзирования образуются при искажении вблизи массивных тел лучей света от далеких источников и, следовательно, могут присутствовать повсюду во Вселенной. Однако события гравитационного линзирования на космических струнах обладают характерным локализованным вдоль струны избыточным распределением, что отличает их от событий обычного гравитационного линзирования.

Таким образом, на всей протяженности струны ожидается формирование цепочек гравитационно-линзовых

пар фоновых по отношению к струне галактик, т.н. «Млечный путь гравитационных линз». Количество таких гравитационно-линзовых событий должно быть избыточно по сравнению с числом обычных гравитационно-линзовых событий и их распределение должно отличаться от соответствующего распределения в обычном случае.

В качестве области поиска таких цепочек линзированных галактик и исследования характеристик их распределений использовались данные обработки карт анизотропии реликтового излучения, в которых был найден кандидат в космическую струну — объект CSc–1, протяженностью от $(\alpha = 11:29:03; \delta = +15:23:37)$ до $(\alpha = 10:57:47; \delta = +25:03:51)$ [13]. В рамках исследования данного протяженного объекта проводилось статистическое сравнение количества гравитационно-линзовых пар галактик в полях, заведомо не содержащих струн, с количеством аналогичных пар в поле этого объекта.

Сравнительный анализ плотности распределения пар линзированных галактик проводился путем доказательства неоднородности двух наборов статистических данных.

14.3 Исходные наблюдательные данные и формирование выборок для статистического анализа

Кандидаты в гравитационно-линзовые события были найдены по фотометрическому каталогу галактик DR12 SDSS [13].

Было сформировано две выборки (см. табл. 35). Каждый элемент 1-й выборки — это количество пар галактик, расположенных в площадке 1 кв. град. в поле CSc–1 (каждая пара галактик находится не далее чем на 0,5

град. от предполагаемого прохождения струны в картинной плоскости; такое расстояние обусловлено качеством исходных данных по анизотропии реликтового излучения, по которым был найден объект CSc–1). Каждый элемент 2-й выборки (контрольная выборка) — это количество пар галактик, расположенных в площадке 1 кв. град. в поле, заведомо не принадлежащем струне (в поле, где по данным анизотропии реликтового излучения струны не обнаружены). Дробность значений количества пар обусловлена нормировкой количества на 1 кв. град.

Таблица 35

Исходные данные по сравнению распределений пар галактик в поле предполагаемой струны и вне этого поля.

N	Поле со струной	Контрольное поле	N	Поле со струной	Контрольное поле
1	52,61	20,06	17	23,23	11,01
2	32,92	16,05	18	27,53	14,44
3	37,38	23,10	19	27,64	16,87
4	26,42	30,09	20	14,91	32,72
5	26,48	27,07	21	21,37	13,98
6	24,31	32,23	22	32,19	23,35
7	13,23	31,13	23	30,11	28,89
8	25,09	20,08	24	17,38	14,23
9	30,21	15,05	25	10,87	11,74
10	10,73	10,75	26	26,21	25,88
11	30,13	19,82	27	15,37	18,19
12	36,71	23,22	28	26,50	23,61
13	38,95	30,52	29	51,08	22,85
14	30,41	12,68	30	17,05	
15	21,06	25,64	31	17,92	
16	29,49	18,85			

Обе выборки представляют собой соответствующие плотности пар гравитационно-линзовых галактик в поле CSc–1 и в контрольных полях без струны.

С помощью непараметрических критериев производилась проверка того, что вне зависимости от способа

группировки данных этих выборок и применяемых методов, плотности распределения данных двух выборок существенно различны и, таким образом, выборка вблизи CSc–1 представляется аномальной по сравнению с контрольной выборкой. Этот факт является дополнительным косвенным доказательством того, что объект CSc–1 есть космическая струна.

Пример такой группировки данных: в поле CSc–1 число площадок, в которых значение плотности пар лежит в интервале $(9, 11)$ равно 2, а для контрольного поля число плотности пар в этом интервале равно 3. Эти числа можно рассматривать как случайные величины, характеризующие частоту, с которой в исследуемых полях встречаются гравитационно-линзовые пары. Таким образом, распределения этих величин можно использовать для доказательства неоднородности двух наборов статистических данных. Разбиение на интервалы плотности пар проводилось двумя способами. В первом случае для поля со струной было получено распределение частот, состоящее из 24 случайных величин, для контрольных полей —- состоящее из 13 величин; после перегруппировки наборы частот включали уже 12 и 7 величин для исследуемого и контрольных полей соответственно (см. табл. 36). Очевидно, что эти выборки являются малыми, поэтому проводя сравнительный анализ параметров распределений, необходимо рассмотреть несколько критериев.

14.4 Статистическая обработка данных

14.4.1 Обоснование использования непараметрических критериев

Для того, чтобы с уверенностью утверждать, что избыток событий гравитационного линзирования в исследу-

Таблица 36

Данные, перегруппированные по интервалам плотности числа пар гравитационно-линзовых объектов в исследуемом и контрольных полях.

Первая перегруппировка			Вторая перегруппировка		
Интервал	Поле со струной ($n_1 = 24$)	Контрольное поле ($n_2 = 13$)	Интервал	Поле со струной ($m_1 = 12$)	Контрольное поле ($m_2 = 7$)
9	0	0	11	2	3
11	2	3	13	3	5
13	1	2	17	3	5
15	2	3	21	3	7
17	3	2	25	8	3
19	0	3	29	5	4
21	2	2	33	2	2
23	1	5	37	3	
25	2	2	41	0	
27	6	1	45	0	
29	1	1	49	1	
31	4	3	53	1	
33	2	2			
35	0				
37	2				
39	1				
41	0				
43	0				
45	0				
47	0				
49	0				
51	1				
53	1				
55	0				

емом поле CSc–1 вызван наличием в этом поле космической струны, необходимо провести анализ наблюдательных данных с помощью методов математической статистики. Это можно сделать, сравнив параметры распределений, характеризующих плотность числа пар линзированных галактик в исследуемом и контрольных полях.

Основная трудность при решении подобных задач заключается в том, что законы распределения вероятностей для рассматриваемых случайных величин заранее не известны. Кроме того, нет возможности увеличить объемы выборок для улучшения точности анализа. Это означает, что применение стандартных параметрических критериев (Стьюдента, Фишера и др.) может привести к малоправдоподобным выводам. Поэтому в данном случае целесообразно использовать свободные от распределения критерии однородности.

В общем случае, непараметрические критерии делятся на критерии сдвига и масштаба (любое распределение может быть описано с помощью параметров положения и масштаба, на сопоставлении которых и основаны указанные выше критерии).

Цель исследования — доказательство несовпадения двух распределений случайных величин. Таким образом, нужно опровергнуть хотя бы одно из предположений: о равенстве параметров положения или о равенстве параметров масштаба.

Будем исследовать сдвиг, определяемый разностью параметров положения, характеризующих центры группирования случайных величин в исследуемых распределениях (к критериям масштаба необходимо будет обратиться только в том случае, если предположение о наличии сдвига будет отклонено).

Для контроля правильности применения критериев в рамках данной задачи, используемые критерии тести-

руются на искусственно сгенерированных (синтезированных) выборках (см. табл. 40), распределения случайных величин в которых заведомо совпадают. Это позволяет удостовериться в адекватности применяемых методов, и, следовательно, в достоверности сделанных на их основе выводов.

14.4.2 Ранговые критерии сдвига

По исходным данным осущетвляется присвоение рангов для объединенных выборок для первой и второй перегруппировки данных (см. табл. 37–38).

Таблица 37

Присвоение рангов элементам для первой перегруппировки данных ($n_1 = 24$, $n_2 = 13$).

№ группы	Частота	Ранг	№ группы	Частота	Ранг	№ группы	Частота	Ранг
1	0	5,5	1	1	14,5	2	2	24
1	0	5,5	1	1	14,5	2	2	24
1	0	5,5	1	1	14,5	2	2	24
1	0	5,5	2	1	14,5	1	3	32
1	0	5,5	2	1	14,5	2	3	32
1	0	5,5	1	2	24	2	3	32
1	0	5,5	1	2	24	2	3	32
1	0	5,5	1	2	24	2	3	32
1	0	5,5	1	2	24	1	4	35
2	0	5,5	1	2	24	2	5	36
1	1	14,5	1	2	24	1	6	37
1	1	14,5	2	2	24			
1	1	14,5	2	2	24			

Ранговых критериев сдвига существует много, но есть три условия, которые существенно ограничивают выбор: наличие повторяющихся значений в наборах случайных величин, неравенство объемов выборок и малый объем выборок. Учитывая вышеперечисленные особенности,

Таблица 38

Присвоение рангов элементам для второй перегруппировки данных ($m_1 = 24$, $m_2 = 13$).

N группы	Частота	Ранг	N группы	Частота	Ранг
1	0	1,5	1	3	10,5
1	0	1,5	2	3	10,5
1	1	3,5	2	3	10,5
1	1	3,5	2	4	14
1	2	6	1	5	16
1	2	6	2	5	16
2	2	6	2	5	16
1	3	10,5	2	7	18
1	3	10,5	1	8	19
1	3	10,5			

можно выделить только три критерия, применение которых в рамках поставленной задачи будет обоснованным: быстрый ранговый критерий, критерий ван дер Вардена и критерий Манна–Уитни–Вилкоксона, частным случаем которого является аппроксимация Имана.

14.4.3 Быстрый ранговый критерий

Быстрый (грубый) ранговый критерий — самый простой с точки зрения вычислительной сложности, его эффективность 86%. Критерий основан на переходе к статистике *d–критерия*, которая может быть аппроксимирована нормальным распределением с нулевым средним. Алгоритм перехода к d–статистике:

- составление общего ранжированного ряда из двух выборок объемами n_1 и n_2;
- присвоение рангов;

- расчет суммарных усредненных рангов для каждой из групп

$$\overline{R}_1 = \frac{1}{n_1}\sum R_i,$$

$$\overline{R}_2 = \frac{1}{n_2}\sum R_j.$$

Тогда d-критерий имеет вид: $d = \overline{R}_1 - \overline{R}_2$ со стандартным отклонением

$$s = s_d = \sqrt{\frac{(n_1+n_2)(n_1+n_2+1)}{12}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}.$$

Проверка гипотезы о неоднородности двух наборов статистических данных осуществляется путем сравнения с u_γ (см. табл. 14): если $|d/s_d| < 1{,}96$, то гипотеза сдвига отклоняется с доверительной вероятностью $\gamma = 0{,}95$. Для данных табл. 36:

- для $n_1 = 24$, $n_2 = 13$:

$$\overline{R}_{n_1} = \frac{1}{24}(5{,}5\cdot 9 + 14{,}5\cdot 6 + 24\cdot 6 + 32 + 35 + 37) \approx 16{,}021;$$

$$\overline{R}_{n_2} = \frac{1}{13}(5{,}5 + 14{,}5\cdot 2 + 24\cdot 5 + 32\cdot 4 + 36) = 24{,}5;$$

$$d = 24{,}5 - 16{,}021 = 8{,}48;$$

$$s_d = \sqrt{\frac{(24+13)(24+13+1)}{12}\left(\frac{1}{24}+\frac{1}{13}\right)} \approx 3{,}73;$$

$$\left|\frac{d}{s_d}\right| = \left|\frac{8{,}48}{3{,}73}\right| \approx 2{,}27 > 1{,}96.$$

- для $m_1 = 12$, $m_2 = 7$:

$$\overline{R}_{m_1} = \frac{1}{12}(1{,}5\cdot 2+3{,}5\cdot 2+6\cdot 2+10{,}5\cdot 4+16+19) \approx 8{,}25;$$

$$\overline{R}_{m_2} = \frac{1}{7}(6+10{,}5\cdot 2+14+16\cdot 2+18) = 13;$$

$$d' = 13-8{,}25 = 4{,}75;$$

$$s'_d = \sqrt{\frac{(12+7)(12+7+1)}{12}\left(\frac{1}{12}+\frac{1}{7}\right)} \approx 2{,}68;$$

$$\left|\frac{d'}{s'_d}\right| = \left|\frac{4{,}75}{2{,}68}\right| \approx 1{,}77 < 1{,}96.$$

Предположение о сдвиге, полученное быстрым ранговым критерием, подтверждается только для большой выборки, поэтому требуются дополнительные исследования другими более эффективными методами.

14.4.4 Критерий ван дер Вардена

Критерий ван дер Вардена различен для выборок средних и малых объемов. Для второй группы данных табл. 36 суммарным объемом $m_1 + m_2 = 19$ статистика ван дер Вардена имеет вид

$$X_m = \sum_{j=1}^{m_2} u_{\Upsilon_j}.$$

Суммирование можно вести и относительно m_1, результат не меняется. Величина $u_{\gamma j}$ вычисляется по приближенной формуле

$$u_{\Upsilon_j} \approx 4{,}91\left[\Upsilon_j^{0{,}14} - (1-\Upsilon_j)^{0{,}14}\right],$$

где

$$\gamma_j = \frac{R_j}{m_1 + m_2 + 1} \ (j = 1, 2, \ldots, m_2).$$

Для второй перегруппировки данных табл. 36:

$$\gamma_1 = \frac{6}{12 + 7 + 1} = 0{,}3;$$

$$u_{\gamma_1} = 4{,}91 \left[0{,}3^{0{,}14} - (1 - 0{,}3)^{0{,}14}\right] \approx -0{,}52;$$

$$\gamma_2 = \frac{10{,}5}{20}; \ \ u_{\gamma_2} \approx 0{,}06;$$

$$\gamma_3 = \frac{14}{20}; \ \ u_{\gamma_3} \approx 0{,}52;$$

$$\gamma_4 = \frac{16}{20}; \ \ u_{\gamma_4} \approx 0{,}84;$$

$$\gamma_5 = \frac{18}{20}; \ \ u_{\gamma_5} \approx 1{,}28;$$

$$X_m = -0{,}52 + 0{,}06 \cdot 2 + 0{,}52 + 0{,}84 \cdot 2 + 1{,}28 = 3{,}08,$$

что меньше критического значения соответствующей статистики 3,62 ($\gamma = 0{,}95; m_1 = 12; m_2 = 7$) [6]. Предположение о сдвиге отклоняется.

Для первой перегруппировки данных табл. 36

$$n = n_1 + n_2 = 37$$

статистика ван дер Вардена аппроксимируется нормальным распределением с нулевым средним и стандартным отклонением

$$s = \sqrt{\frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 1)} \sum_{p=1}^{n_1+n_2} u_{\gamma_p}^2},$$

где

$$\gamma_p = \frac{R_p}{n_1 + n_2 + 1} \ (p = 1, 2, \ldots, n).$$

Соответствующие расчеты дают:

$$X_n = \sum_{j=1}^{n_2=13} u_{Y_j} = 5{,}64;$$

$$s = \left[\frac{24 \cdot 13}{(24+13)(24+13-1)} \sum_{p=1}^{n=37} u^2_{\frac{R_p}{24+7+1}} \right]^{0,5} = 2{,}49.$$

Как и в случае быстрого рангового критерия, величину $|X_n/s|$ требуется сравнить с $u_{0,95}$:

$$\left| \frac{X_n}{s} \right| = \left| \frac{5{,}64}{2{,}49} \right| \approx 2{,}26 > 1{,}96.$$

Таким образом, для первой группы наблюдательных данных табл. 36 предположение о сдвиге справедливо. Однако по отношению к малым выборкам требуется дополнительное исследование, поскольку метод обладает высокой эффективностью (его мощность равна мощности t-критерия Стьюдента) только для больших выборок.

14.4.5 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

$$U = \sum_{i,=,1}^{k_1} \sum_{j,=1}^{k_2} h_{ij}; \quad h_{ij} = \begin{cases} 1, & \text{если } x_i < y_j \\ 0, & \text{если } x_i > y_j. \end{cases}$$

Здесь k_1, k_2 — объемы двух выборок в рассматриваемой группе; $i = 1, 2, \ldots, k_1; j = 1, 2, \ldots, k_2$.

Для вычисления U–статистики необходимо вычислить количество элементов первой выборки, не превосходящих по своему значению случайные величины из второй выборки. При этом не важно, относительно какой из

выборок ведется суммирование. Мы сопоставили величины, соответствующие CSc–1 и контрольному полю, и получили следующие значения U–статистики для первой и второй перегруппировок данных: $U_n = 57; U_m = 15$.

В случае малых выборок наличие сдвига признается, если найденная величина U не входит в числовой интервал, определяемый критическими значениями U_1 и U_2 [6]. Для $m_1 = 12$; $m_2 = 7$ получаем $U_1 = 18; U_2 = 63$.

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

$$W = \frac{R - \frac{n_2(n_1+n_2+1)}{2}}{g}.$$

Величина g в знаменателе данной статистики учитывает совпадающие элементы в выборках, благодаря чему полученное распределение может быть аппроксимировано нормальным. Значение g можно вычислить как

$$g = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}} \times$$

$$\times \sqrt{1 - \frac{\sum_{s=1}^{q} t_s(t_s^2 - 1)}{(n_1+n_2)(n_1+n_2+1)(n_1+n_2-1)}},$$

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.

Таким образом, для $n_1 = 24$ и $n_2 = 13$ получаем

$$R = 24 \cdot 13 + \frac{13(13+1)}{2} - 57 = 346.$$

Величина $q = 4$. Следовательно,

$$t_1(0) = 10;\ t_2(1) = 8;\ t_3(2) = 11;\ t_4(3) = 5.$$

В скобках при t_s указаны численные значения совпадающих элементов в наборе данных.

$$g = \left[\frac{24 \cdot 13 \cdot (n_1 + n_2 + 1)}{12} \Big(1 - \right.$$

$$\left. - \frac{\sum\limits_{s=1}^{4} t_s(t_s^2 - 1)}{(24+13) \cdot (24+13+1) \cdot (24+13-1)} \Big) \right]^{0,5} = 30{,}51;$$

$$W = \frac{R - \dfrac{13 \cdot (24+13+1)}{2}}{g} \approx 3{,}25 > 1{,}96.$$

Таким образом, гипотеза сдвига принимается как для первой, так и для второй (малой) группировки исходных данных.

14.4.6 Аппроксимация Имана

Одним из наиболее точных наряду с рассмотренными непараметрическими методами является метод *аппроксимации Имана*. Эффективность метода 95%.

J–статистика этого метода строится на основе W-статистики Вилкоксона:

$$J = \frac{W}{2} \left[1 + \left(\frac{n_1 + n_2 - 2}{n_1 + n_2 - 1 - W^2} \right)^{0,5} \right].$$

Полученное значение сравнивается с критическим:

$$J_\gamma = (u_\gamma + t_\gamma)/2.$$

Здесь u_γ и t_γ — γ–квантили нормального распределения и распределения Стьюдента с $r = n_1 + n_2 - 2$ степенями свободы соответственно.

Расчетное значение для второй перегруппировки данных:

$$J = \frac{3{,}25}{2}\left[1 + \left(\frac{24 + 13 - 2}{24 + 13 - 1 - 3{,}25^2}\right)^{0{,}5}\right] = 3{,}53 > 2{,}00,$$

что свидетельствует о подтверждении предположения о сдвиге с доверительной вероятностью $\gamma = 0{,}95$.

14.4.7 Результаты статистической обработки негруппированных исходных данных

Приведем результаты статистической обработки исходных наблюдательных данных ($n_1 = 31$, $n_2 = 29$, см. табл. 35).

Со стороны рассмотренных непараметрических критериев никаких теоретических ограничений сверху на объемы выборок нет, поэтому можно проанализировать их с помощью этих критериев. Рангами рассматриваемых величин будут их порядковые номера в общем ранжированном ряду, потому что совпадающих значений среди элементов выборок нет.

В табл. 39 приводятся результаты вычислений для исходных данных и сводка всех расчетных величин, полученных при первой и второй группировках исходных данных.

Таблица 39

Итоговые результаты для исследуемых выборок.

Непараметрический критерий сдвига	$N_1 = 31$, $N_2 = 29$	$n_1 = 24$, $n_2 = 13$	Критическое значение	$m_1 = 12$, $m_2 = 7$	Критическое значение
Быстрый ранговый	2,17	2,27	1,96	1,77	1,96
Ван дер Вардена	2,12	2,26	1,96	3,02	3,62
Манна-Уитни-Вилкоксона	2,18	3,25	1,96	15	18 <...< 63
Аппроксимация Имана	2,22	3,53	2.00	-	-

14.4.8 Тестирование используемых методов на синтезированных выборках

Чтобы продемонстрировать справедливость полученных результатов, были протестированы все использованные ранговые критерии сдвига на заведомо однородных выборках, представленных в табл. 40. Были взяты выборки того же объема и структуры, что и рассмотренные ранее. Так, например, в однородных выборках с $n_2 = 24$, $m_2 = 13$ имеются повторяющиеся числа, а диапазон значений — целочисленный, от 0 до 6. Как можно видеть из табл. 41, использованные для анализа критерии однозначно подтверждают одинаковость распределения случайных величин в тестовых выборках, что указывает на устойчивость работы критериев.

14.4.9 Выводы

Была решена задача статистического сравнении распределения гравитационно-линзовых пар галактик в по-

Таблица 40

Синтезированные 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	16,17	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	15,15	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			
28,90	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					

Таблица 41

Результаты анализа однородных синтезированных 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	-	-

лях, заведомо не содержащих космических струн (контрольное поле), с распределением аналогичных пар в поле CSc–1, где по данным анизотропии реликтового излучения [13] расположен кандидат в космическую струну. В случае действительного присутствия космической струны в соответствующем поле должен наблюдаться не только статистический избыток гравитационно-линзовых пар (который был получен в [13]), но и сама плотность распределения гравитационно-линзовых пар должна отличаться от соответствующей плотности распределения классических событий в не содержащих струн полях.

Сравнительный анализ плотности распределения пар линзированных галактик проводился путем доказательства неоднородности двух наборов статистических данных. Выбранные для этой цели методы математической статистики, а именно непараметрические ранговые критерии сдвига позволили работать с малыми выборками, плотности распределения которых заранее не известны.

Также обсуждалась правильность работы используемых критериев на основе статистической обработки синтезированных выборок с заранее известными параметрами.

Полученные результаты указывают на статистическое различие распределений гравитационно-линзовых пар в поле CSc–1 по сравнению с контрольным полем. Выявленное различие служит дополнительным аргументом в пользу наличия в исследуемом поле космической струны.

Здесь отметим, что если бы исследованный сдвиг оказался статистически незначимым, то аналогичные исследования следовало бы проделать для параметров масштаба сравниваемых выборок с использованием тех же критериев и их модификаций, которые подробно описаны, например, в [6]. Выборки признавались бы статистически идентичными (полученными в одинаковых физических условиях, что исключило бы гипотезу космической струны, влияющей на одну из выборок) только тогда, когда и параметры положения, и параметры масштаба оказались бы статистически равными.

Приложение. Понятие фактора Байеса

Фактор Байеса — это число, которое есть апостериорная вероятность предполагаемой (нулевой) гипотезы (в простейшем случае априорная вероятность этой нулевой гипотезы есть 0,5). Фактор Байеса предлагает способ численной оценки свидетельств в пользу нулевой гипотезы. Носит общий характер и не требует дополнительных исследований.

Есть несколько техник расчета фактора Байеса. Например, асимптотическая аппроксимация, которая вычисляется с помощью ММП. Также используется критерий Шварца (один из информационных критериев). В случае, когда рассматривается модель с ошибками, распределенными нормально (для нормальной линейной регрессии) критерий Шварца имеет вид:

$$CS = \ln \hat{\sigma}^2 + \frac{k \cdot \ln n}{n},$$

где $\hat{\sigma}^2 = \bar{\varepsilon}^2/n$ есть оценка дисперсии остатков, вычисленная по выборке; k — число параметров. Отметим, что логарифмическая функция правдоподобия для нормальной линейной регрессии есть

$$\ln L = -\frac{n}{2} \cdot \left(1 + \ln 2\pi + \ln \hat{\sigma}^2\right).$$

В общем случае критерий Шварца есть:

$$CS = k \cdot \ln n - 2 \ln L.$$

Критерий Шварца дает грубую аппроксимацию величины логарифма фактора Байеса, легкую в использовании и не требующую оценки апостериорных вероятностей.

Рассмотрим событие A (которое есть набор неких данных D) и две гипотезы H_1 и H_2.

Формула Байеса ($k = 1, 2$):

$$P(H_k|D) = \frac{P(H_k) \cdot P(D|H_k)}{P(H_1) \cdot P(D|H_1) + P(H_2) \cdot P(D|H_2)}.$$

Можно переписать эту формулу следующим образом:

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(H_1) \cdot P(D|H_1)}{P(H_2) \cdot P(D|H_2)}.$$

Фактор Байеса по определению есть

$$B \equiv \frac{P(D|H_1)}{P(D|H_2)}.$$

Удвоенный логарифм от этой величины называется *логарифмический фактор Байеса.*

Другими словами, «апостериорная вероятность = фактор Байеса × априорная вероятность» (послеопытная вероятность, т.е. полученная уже при наличии данных D, равна произведению этого фактора на доопытную вероятность). Фактор Байеса — отношение апостериорной вероятности осуществления гипотезы H_1 к ее априорной вероятности, независимо от величины априорной вероятности.

Если гипотезы H_1 и H_2 априорно равновероятны, т.е. $P(H_1) = P(H_2) = 0{,}5$, то фактор Байеса равен апостериорной вероятности в пользу гипотезы H_1. Если обе гипотезы не содержат свободных параметров, то этот фактор есть отношение правдоподобия.

В более сложных случаях, когда присутствуют неизвестные параметры, например, $(\vartheta_1, \vartheta_2)$ в одной или обеих гипотезах, фактор есть

$$B \equiv \frac{P(D|H_1)}{P(D|H_2)} = \frac{\int f(D|\vartheta_1, H_1) \cdot \pi(\vartheta_1|H_1) d\vartheta_1}{\int f(D|\vartheta_2, H_2) \cdot \pi(\vartheta_2|H_2) d\vartheta_2}.$$

Здесь $\pi(\vartheta_1|H_1)$ и $\pi(\vartheta_2|H_2)$ есть априорные (доопытные) вероятности значений параметров в условиях соответствующих гипотез H_1, H_2. Величины $f(D|\vartheta_1, H_1)$ и $f(D|\vartheta_2, H_2)$ есть плотности вероятности D в условиях соответствующих параметров ϑ_1, ϑ_2 и гипотез H_1, H_2. Другими словами, величины $f(D|\vartheta_1, H_1)$ и $f(D|\vartheta_2, H_2)$ есть функции правдоподобия параметров ϑ_1, ϑ_2.

В еще более общем случае параметры являются векторами. В выражении для фактора Байеса числитель и знаменатель называются «маргинальными» (или «интегральными») функциями правдоподобия. Параметры ϑ_1, ϑ_2 могут быть исключены из выражения для фактора Байеса не только путем интегрирования, но и путем процедуры максимизации по ним.

Интеграл вида

$$I = \int f(D|\vartheta, H) \cdot \pi(\vartheta|H) d\vartheta$$

может быть вычислен аналитически для экспоненциального распределения и родственных ему, а также для нормальных линейных моделей.

Аппроксимацию этого интеграла можно получить, используя следующее предположение (аппроксимация Лапласа). Апостериорная плотность вероятности, которая пропорциональна $f(D|\vartheta, H) \cdot \pi(\vartheta|H)$, обладает выраженным пиком вблизи некого значения параметра $(\tilde{\vartheta})$, которое есть апостериорная мода. Разлагая функцию $\log f(D|\vartheta, H) \cdot \pi(\vartheta|H)$ в ряд по параметру ϑ в окрестности $\tilde{\vartheta}$ и потом снова записывая в виде экспоненты, получаем разложения для $f(D|\vartheta, H) \cdot \pi(\vartheta|H)$, которое имеет форму нормальной плотности вероятности со средним $\tilde{\vartheta}$ и ковариационной матрицей

$$\tilde{\Sigma} = (-D^2 \log f(D|\tilde{\vartheta}, H) \cdot \pi(\tilde{\vartheta}|H))^{-1}.$$

Здесь $D^2 \log f(D|\tilde{\vartheta}, H) \cdot \pi(\tilde{\vartheta}|H)$ — это матрица Гессе вторых производных. Интегрирование полученной аппроксимации дает оценку интеграла I (d — размерность вектора параметров):

$$\hat{I} = \frac{1}{\sqrt{(2\pi)^d}} |\tilde{\Sigma}|^{1/2} f(D|\tilde{\vartheta}, H) \cdot \pi(\tilde{\vartheta}|H),$$

$$I = \hat{I}\big(1 + O(n^{-1})\big), n \to \infty.$$

Модификация аппроксимации Лапласа есть интеграл

$$\hat{I}_{mod} = \frac{1}{\sqrt{(2\pi)^d}} |\hat{\Sigma}|^{1/2} f(D|\hat{\vartheta}, H) \cdot \pi(\hat{\vartheta}|H).$$

Здесь $\hat{\Sigma}^{-1}$ — наблюдаемая информационная матрица Фишера, т.е. обратная матрица Гессе для логарифмической функции правдоподобия для оценки $\hat{\vartheta}$. В этом случае:

$$I = \hat{I}\big(1 + O(n^{-1/2})\big).$$

Если аппроксимировать величину $\tilde{\vartheta}$ величиной $\hat{\vartheta}$ (одношаговым методом Ньютона) и подставить в выражение

$$\hat{I} = \frac{1}{\sqrt{(2\pi)^d}} |\tilde{\Sigma}|^{1/2} f(D|\tilde{\vartheta}, H) \cdot \pi(\tilde{\vartheta}|H),$$

то фактор Байеса примет вид

$$2 \log B =$$

$$= \frac{2 \log P(D|H_1)}{\log P(D|H_2)} = \frac{2 \int \log f(D|\vartheta_1, H_1) \cdot \pi(\vartheta_1|H_1) d\vartheta_1}{\int \log f(D|\vartheta_2, H_2) \cdot \pi(\vartheta_2|H_2) d\vartheta_2} \approx$$

$$\approx \Lambda + (E_1 - E_2),$$

где

$$\Lambda = 2 \log f(D|\hat{\vartheta}_1, H_1) - 2 \log f(D|\hat{\vartheta}_2, H_2),$$

$$E_1 =$$

$$= 2\log\pi(\hat{\vartheta}_1|H_1) + \left\{\frac{d}{d\vartheta_1}\log\pi(\vartheta_1|H_1)\right\}\cdot\left(\hat{\Sigma}_1^{-1} + \left(D\left[\frac{\vartheta_1}{H_1}\right]\right)^{-1}\right)^{-1}\cdot$$

$$\cdot\left\{2 - \hat{\Sigma}_1^{-1}\cdot\left(\hat{\Sigma}_1^{-1} + \left(D\left[\frac{\vartheta_1}{H_1}\right]\right)^{-1}\right\}^{-1}\cdot\left\{\frac{d}{d\vartheta_1}\log\pi(\vartheta_1|H_1)\right\} -$$

$$- \log\left|\hat{\Sigma}_1^{-1} + \left(D\left[\frac{\vartheta_1}{H_1}\right]\right)^{-1}\right| + \log 2\pi.$$

Для E_2 выражение аналогично. Производные берутся в точках $\hat{\vartheta}_1$ и $\hat{\vartheta}_2$ соответственно.

При вычислении фактора Байеса можно воспользоваться критерием Шварца, обойти необходимость вычисления $\pi(\vartheta_1|H_1)$, $\pi(\vartheta_2|H_2)$ введя следующую величину (для простоты считаем, что имеется всего два параметра, один скалярный, а другой размерностью на единицу больше, т.н. модель «вложенных» гипотез):

$$S = \log f(D|\hat{\vartheta}_1, H_1) - \log f(D|\hat{\vartheta}_2, H_2) - \frac{1}{2}\log n;$$

$$\frac{S - \log B}{\log B} \to 0.$$

Правдоподобность фактора Байеса можно грубо оценить с использованием критерия Шварца и с помощью критерия Стьюдента, для большого количества данных. Это можно сделать следующим образом. В частном случае в выражении для Λ

$$\Lambda = 2\log f(D|\hat{\vartheta}_1, H_1) - 2\log f(D|\hat{\vartheta}_2, H_2)$$

оценки параметров имеют нормальный закон распределения (тогда функция f — это нормальная плотность

распределения). Тогда Λ пропорциональна квадрату нормальной статистики и, используя формулу Шварца, получаем:

$$2\log B \approx u^2 - \log n.$$

В более общем случае величина

$$\frac{\hat{\vartheta}_1 - \hat{\vartheta}_2}{s.d./\sqrt{n}} \sim t$$

и тогда

$$2\log B \approx t^2 - \log n$$

при дополнительном условии, $\hat{\vartheta}_1 - \hat{\vartheta}_2 \sim O\left(n^{-1/2}\right)$.

Однако поскольку критерий Шварца верен для больших выборок, то можно пользоваться формулой с нормальной статистикой вместо статистики Стьюдента.

Значимость логарифмического фактора Байеса традиционно определяется согласно нижеследующей таблице (см. табл. 42).

Таблица 42

Значимость логарифмического фактора Байеса.

$2\log B$	B	Весомость доказательства
$[0;1]$	$[1;3{,}2]$	Слабая
$[1;2]$	$[3{,}2;10]$	Положительная
$[2;4]$	$[10;100]$	Сильная

Таким образом, используя критерий Стьюдента (который в условиях $n > 30$ рекомендуется заменять на нормальный) можно получить количественные характеристики весомости доказательства фактора Байеса. Например, для $n = 45$ и $2\log B = 0{,}74$ (слабая весомость

доказательства) на основе вышеприведенных формул величина t^2 есть 2,39, что соответствует уровню достоверности 93,54%. Для $n = 45$ и $2 \log B = 1{,}28$ (положительная весомость доказательства) t^2 есть 2,93, что соответствует уровню достоверности 95,31%. Наконец, для $n = 45$ и $2 \log B = 5{,}4$ (сильная весомость доказательства) t^2 есть 7,05, что соответствует уровню достоверности 99,45% (Фактор Байеса рассчитан для данных гравитационно-волнового эксперимента NANOGrav, 2020).

Использованная литература

[1] Агекян Т.А. Основы теории ошибок для астрономов и физиков. М.: Наука, 1972.

[2] ван дер Варден Б.Л. Математическая статистика. М.: Изд-во иностр. литер., 1960.

[3] Демидович Б.П., Марон И.А., Шувалов Э.З. Численные методы анализа. Приближение дифференциальные и интегральные уравнения. М.: Наука, 1967.

[4] Деммель Дж. Вычислительная линейная алгебра. Теория и приложения. М.: Мир, 2001.

[5] Ильин В.А., Позняк Э.Г. Линейная алгебра. М.: Наука; Физматлит, 1999.

[6] Кобзарь А.И. Прикладная математическая статистика. М.: Физматлит, 2006 (и более поздние издания).

[7] Кочетков Е.С., Осокин А.В. Случайные события. М.: МАИ, 2000.

[8] Мардиа К. Статистический анализ угловых наблюдений. М.: Наука, 1978.

[9] Монсик В.Б., Скрынников А.А. Вероятность и статистика. М.: Бином. Лаборатория знаний, 2011.

[10] Сборник задач по математике для втузов: в 4 ч. / под ред. А.В. Ефимова, А.Н. Поспелова. М.: Физматлит, 2003.

[11] Худсон Д. Статистика для физиков. М.: Мир, 1970.

[12] Щиголев Б.М. Математическая обработка наблюдений. М., 1969.

[13] Sazhina O.S., Scognamiglio D., Sazhin M.V. et al. Optical analysis of a CMB cosmic string candidates // MNRAS. 2019. Vol. 485, N 2. P.1876-1885.

[14] Wasserman L. All of Statistics. A Concise Course in Statistical Inference. New York: Springer, 2004.

Рекомендуемая литература

Теория вероятностей и математическая статистика

[15] Вентцель Е.С. Теория вероятностей. М.: Наука, 1969.

[16] Гмурман В.Е. Теория вероятностей и математическая статистика. М.: Юрайт, 2018.

[17] Гнеденко Б.В. Курс теории вероятностей. М.: Наука, 1988.

[18] Дрейнер Н., Смит Г. Прикладной регрессионный анализ. М.: Диалетика, 2016.

[19] Колмогоров А.Н. Основные понятия теории вероятностей. М.: Наука, 1974.

[20] Крамер Г. Математические методы статистики. М.: Наука, 1975.

[21] Линник Ю.В. Метод наименьших квадратов и основы теории обработки наблюдений. М.: Физматгиз, 1962.

[22] Лоэв М. Теория вероятностей. М.: Изд-во иностр. литер., 1962.

[23] Пугачев В.С. Введение в теорию вероятностей. М.: Наука, 1968.

[24] Пугачев В.С. Теория вероятностей и математическая статистика. М.: Наука, 1979.

[25] Феллер В. Введение в теорию вероятностей и ее приложений: в 2 т. . М.: Мир, 1984.

Линейная алгебра

[26] Гантмахер Ф.Р. Теория матриц. М.: Физматлит, 2010.

[27] Гельфанд И.М. Лекции по линейной алгебре. М.: Наука, 1971.

[28] Кострикин А.И. Введение в алгебру. Ч. II: Линейная алгебра. М.: Наука, 2000.

Численные методы

[29] Бахвалов И.С., Жидков Н.П., Кобельков Г.М. Численные методы. М.: Бином, 2020.

[30] Калиткин Н.Н. Численные методы. СПб.: БХВ-Петербург, 2011.

Предметный указатель