1 Описательная статистика

1.1 Выборка, эмпирическое распределение, гистограмма (л.1)

Главной задачей математической статистики является разработка методов построения научно-обоснованных выводов о массовых явлениях и процессах на основе данных наблюдений и экспериментов. Эти выводы касаются параметов, видов распределения и других свойств случайных величин по конечной совокупности наблюдений над ними — выборке.

Выборка понимается следующим образом. Пусть случайная величина ξ наблюдается в случайном эксперименте, который повторяется n раз при одних и тех же условиях. Этот составной эксперимент связан со случайным вектором $(\xi_1, \ldots, \xi_n)$, где случайная величина ξ_j соответствует j-му эксперименту. В биостатистике с понятием эксперимента соотносится понятие индивида. Очевидно, что компоненты ξ_j, $j = 1, 2, \ldots, n$, — независимые в совокупности и распределенные по тому же закону, что и случайная величина ξ.

Закон распределения случайной величины ξ называется *законом распределения генеральной совокупности*, а случайный вектор $(\xi_1, \ldots, \xi_n)$ выборочным вектором. Реализация выборочного вектора называется *выборкой* $(x_1, \ldots, x_n)$ объема n.

Аналогично определяется выборка в случае нескольких случайных величин. Обычно p-мерную выборку представляют в виде следующей таблицы.

	ξ_1	ξ_2	$\ldots$	ξ_p
ω_1	x_{11}	x_{12}	$\ldots$	x_{1p}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
ω_n	x_{n1}	x_{n2}	$\ldots$	x_{np}

Через ω_j, $j = 1, \ldots, n$, обозначаются *индивиды*, через ξ_i, $i = 1, \ldots, p$, *признаки*, через ξ_{ji} *варианты* или значения признака ξ_i у индивида ω_j.

Например, в табл.1.1 представлены данные об оценках за контрольную работу в разных группах студентов географического факультета.

Таблица 1: Средние оценки в группе за контрольную работу по статистике, (географический факультет)

группа	2005г	2006г
политическая география	2.59	2.55
геоэкология	2.32	1.73
экологическая безопасность	2.43	2.25
картография	3.2	2.25
биогеография	2.25	1.33
страноведение	3.08	2.05
геоморфология	2.88	1.71
экономическая география	2.96	1.87
физическая география	2.50	1.89

Признаки подразделяются на *количественные, порядковые и качественные или категориальные*. К количественным признакам относятся те, которые можно измерить в определенном масштабе (оценки за контрольную работу), к порядковым те, которые измерить нельзя, но можно упорядочить, например, при построении по росту. Тяжесть заболевания с градациями: стабильное состояние, средней тяжести и тяжелое — относится к порядковому признаку. Основное отличие качественных признаков заключается в том, что их градации можно менять местами. Например, цвет глаз, тип операции, назначенный для лечения препарат и так далее.

Пусть выборка $(x_1, \ldots, x_n)$ содержит k различных градаций $z_1, \ldots, z_n$ признака ξ, причем градация z_i встречается n_i раз, $\sum_{i=1}^{k} n_i = n$. *Статистическим рядом* называется последовательность пар (z_i, n_i), $i = 1, \ldots, k$.

В случае количественных или порядковых признаков применяют *вариационный ряд*, под которым понимают упорядочен-

ную выборку

$$x^{(1)} \leq \ldots \leq x^{(n)}.$$

Разность $x^{(n)} - x^{(1)} = R$ называется *размахом выборки.*

При большом объеме выборки ее элементы объединяются в группы (разряды, карманы и т.п.), представляя результаты в виде *группированного статистического ряда.* Для этого интервал, содержащий все элементы выборки, разбивается на k непересекающихся интервалов. Длина этих интервалов обычно одинакова и равна $b \approx \frac{\omega}{k}$. После этого вычисляются частоты ν_i, равные количеству элементов выборки, попадающих в этот интервал. Через z_i обозначаются середины интервалов группировки, частоты $\frac{\nu_i}{n}$ называются *относительными.*

Определение 1 Пусть $x_1, \ldots, x_n$ выборка из генеральной совокупности с функцией распределения $F(x)$. *Распределением выборки* называется распределение дискретной случайной величины, принимающей значения $x_1, \ldots, x_n$ с вероятностями $1/n$. Соответствующая функция распределения называется выборочной или *эмпирической функцией распределения*

$$F_n(x) = \frac{\mu_n(x)}{n},$$

где $\mu_n(x)$ равно количеству элементов выборки, не больших x.

Теорема 1 *(Гливенко) Для любого* $x \in (-\infty; +\infty)$ *и для любого* $\epsilon > 0$

$$\lim_{n \longrightarrow \infty} P\{|F_n(x) - F(x)| < \epsilon\} = 1.$$

Действительно, если считать „успехом“ событие $x_i \leq x$ с вероятностью $p = P\{x_i \leq x\}$, то $\mu_n(x)$ равно числу успехов в n независимых испытаниях. Необходимое утверждение получаем на основании теоремы Бернулли

$$\lim_{n \longrightarrow \infty} P\left\{\left|\frac{\mu_n(x)}{n} - p\right| < \epsilon\right\} = 1.$$

Таким образом, при каждом x эмпирическая функция $F_n(x)$ сходится по вероятности к $F(x)$ и при большом объеме выборки может служить приближенным значением (оценкой) функции распределения.

Гистограммой частот группированной выборки будем называть кусочно-постоянную функцию, принимающую на интервалах группировки значения n_i/b. Площадь под графиком равна n. *Гистограмма относительных частот* определяется аналогично с площадью под ступенчатым графиком, равной 1. При увеличении объема выборки и уменьшении интералов группировки гистограмма относительных частот является статистическим аналогом плотности распределения генеральной совокупности.

Полигоном частот называется ломаная с вершинами $\left(z_i, \frac{n_i}{b}\right)$, а в случае относительных частот (z_i — середины интервалов группировки) с вершинами $\left(z_i, \frac{n_i}{nb}\right)$.

1.2 Характеристики выборочного распределения

Числовые характеристики выборочного распределения называются выборочными или эмпирическими. *Выборочное среднее* и *выборочная дисперсия* имеют соответственно вид

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Используя выражения для начальных и центральных выборочных моментов

$$a_\nu = \frac{1}{n}\sum_{i=1}^{n} x_i^\nu, \qquad m_\nu = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^\nu,$$

нетрудно показать, что

$$\begin{aligned} m_2 &= a_2 - \bar{x}^2\,, \\ m_3 &= a_3 - 3a_2\bar{x} + 2\bar{x}^3\,, \\ m_4 &= a_4 - 4a_3\bar{x} + 6a_2\bar{x}^2 - 3\bar{x}^4\,. \end{aligned}$$

Выборочной *модой* унимодального распределения является элемент выборки mod, встречающийся с наибольшей частотой. Например, в табл.1.1 $mod = 2.25$ по данным 2005 и 2006 года.

Выборочной *медианой* является число, которое делит вариационный ряд на две части, содержащие одинаковое число элементов.

Если $n = 2k$, то $med = (x^{(k)} + x^{(k+1)})/2$. Если $n = 2k + 1$, то $med = x^{(k+1)}$. По данным 2005 и 2006 года $n = 18$, $x^{(9)} = 2.25, x^{(10)} = 2.32, med = 2.285$.

С понятием функции распределения неразрывно связано понятие P-квантили распределения — такого значения x_P случайной величины ξ, что

$$P\{\xi \leq x_P\} = P. \tag{1}$$

Если nP — не целое число, то выборочной квантилью x_P^* порядка P называется k-й член вариационного ряда, где $k = \lfloor nP \rfloor + 1$. Если $nP = k$, то выборочная квантиль x_P^* может принимать любое значение на интервале $[x^{(k)}, x^{(k+1)})$. Для определенности используют их среднее арифметрическое.

Коэффициент асимметрии определяется как

$$\gamma_1 = \frac{\mu_3}{\mu_2^{\frac{3}{2}}}, \quad \text{выборочный как} \quad g_1 = \frac{m_3}{m_2^{\frac{3}{2}}}.$$

Для вычисления используем соотношение между центральными и начальными моментами: $m_2 = a_2 - \bar{x}^2$ и $m_3 = a_3 - 3a_2\bar{x} + 2\bar{x}^3$.

Предложение 1 *В случае симметричного закона распределения, когда для плотности распределения справедливо*

$$f(\mathbf{E}\xi - x) = f(\mathbf{E}\xi + x),$$

все нечетные моменты равны нулю.

Действительно,

$$\mu_{2k+1} = \int\limits_{-\infty}^{+\infty} (x - \mathbf{E}\xi)^{2k+1} f(x)dx =$$

$$= \int\limits_{-\infty}^{\mathbf{E}\xi} (x - \mathbf{E}\xi)^{2k+1} f(x)dx + \int\limits_{\mathbf{E}\xi}^{+\infty} (x - \mathbf{E}\xi)^{2k+1} f(x)dx$$

В первом интеграле сделаем замену $x - \mathbf{E}\xi = y$, а во втором $x - \mathbf{E}\xi = -y$. Получаем

$$\int\limits_{-\infty}^{0} y^{2k+1} f(y + \mathbf{E}\xi)dy + \int\limits_{0}^{-\infty} (-y)^{2k+1} f(\mathbf{E}\xi - y)(-dy) = 0\,.$$

Эксцесс $\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$ используется в качестве метрики отклонения от нормального закона распределения, при котором эксцесс равен нулю. Выборочный эксцесс вычисляется как $g_2 = \frac{m_4}{m_2^2} - 3$.

Предложение 2 *Пусть* $\xi \sim \mathcal{N}(0, 1)$. *Тогда* $\mu_4 = 3$.

В плотности нормального распределения заменим σ^2 на $\frac{1}{h}$

$$\int\limits_{-\infty}^{+\infty} e^{\frac{-t^2 h}{2}} dt = \sqrt{2\pi} h^{-\frac{1}{2}}$$

и сосчитаем производные по h:

$$\int\limits_{-\infty}^{+\infty} e^{\frac{-t^2 h}{2}} \left(-\frac{t^2}{2}\right) dt = \sqrt{2\pi} \left(-\frac{1}{2}\right) h^{-\frac{3}{2}},$$

$$\int\limits_{-\infty}^{+\infty} e^{\frac{-t^2 h}{2}} \left(-\frac{t^2}{2}\right)^2 dt = \sqrt{2\pi} \left(-\frac{1}{2}\right)\left(-\frac{3}{2}\right) h^{-\frac{5}{2}}.$$

При $h = 1$ из последнего уравнения получаем

$$\int\limits_{-\infty}^{+\infty} e^{\frac{-t^2 h}{2}} \left(\frac{t^4}{4}\right) dt = \sqrt{2\pi} \left(\frac{3}{4}\right), \implies \mu_4 = 3.$$

2 Оценивание характеристик распределения по выборке

2.1 Свойства оценок параметров (л.2)

Пусть имеется выборка $x_1, \ldots, x_n$ наблюдений над случайной величиной ξ с функцией распределения $F(x, \theta)$, зависящей от параметра θ, например, с математическим ожиданием μ или дисперсией σ^2.

Точечной оценкой параметра называется приближенное значение $\hat{\theta}$ этого параметра, вычисленное по выборке.

Очевидно, оценка $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ является значением некоторой функции от элементов выборки, которая называется *статистикой.* Для выяснения свойств оценки ее рассматривают как функцию от случайного вектора $(\xi_1, \ldots, \xi_n)$, одной из реализаций которого является данная выборка $x_1, \ldots, x_n$. Распределение статистики $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ зависит от неизвестного параметра θ, так как от него зависит закон распределения компонент случайного вектора.

2.1.1 Несмещенность

Определение 2 Оценка $\hat{\theta}_n$ называется *несмещенной,* если ее математическое ожидиние совпадает с истинным значением параметра, то есть $\mathbf{E}\hat{\theta}_n = \theta$.

Если $\mathbf{E}\hat{\theta}_n = \theta + b(\theta)$, то $b(\theta)$ называется *смещением.*

В качестве примера несмещенной оценки рассмотрим выборочное среднее как оценку математического ожидания.

$$\mathbf{E}\bar{x} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}x_i = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu\,. \tag{2}$$

Также несмещенными оценками среднего оказываются $\hat{\mu} = x_i$, так как $\mathbf{E}x_i = \mu\,.$

В качестве примера смещенной оценки рассмотрим второй выборочный центральный момент m_2 как оценку дисперсии. Нетрудно показать, что

$$m_2 = a_2 - \bar{x}^2 \,. \tag{3}$$

Действительно,

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \\ = a_2 - 2\bar{x}\bar{x} + \bar{x}^2 = a_2 - \bar{x}^2 \,.$$

Утверждение 1.

$$\mathbf{E}m_2 = \frac{\sigma^2(n-1)}{n}. \tag{4}$$

Введем $y_i = x_i - \mu$ с математическим ожиданием $\mathbf{E}y_i = 0$ и дисперсией $\mathbf{D}y_i = \sigma^2$, тогда $\bar{y} = \bar{x} - \mu$,

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \\ \text{из (41)} \quad = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2 \,.$$

$$\mathbf{E}m_2 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}y_i^2 - \mathbf{E}\bar{y}^2 = \sigma^2 - \frac{1}{n^2}\sum_{i,j}\mathbf{E}y_i y_j = \\ = \sigma^2 - \frac{1}{n^2}\sum_{i}\mathbf{E}y_i^2 = \sigma^2 - \frac{\sigma^2}{n} \,,$$

так как $\mathbf{E}y_i^2 = Dy_i = \sigma^2$, а при $i \neq j$ из-за независимости элементов выборки $\mathbf{E}y_i y_j = \mathbf{E}y_i \mathbf{E}y_j = 0$.

В качестве несмещенной оценки дисперсии рассматривают статистику

$$S^2 = \frac{n}{n-1} m_2 \,. \tag{5}$$

2.1.2 Состоятельность

Определение 3 Оценка $\hat{\theta}(x_1, \ldots, x_n) = \hat{\theta}_n$ является *состоятельной*, если она по вероятности сходится к истинному значению параметра, то есть $\forall \epsilon > 0$

$$\lim_{n \longrightarrow \infty} P\{|\hat{\theta}_n - \theta| < \epsilon\} = 1\,.$$

Теорема 2 *Если оценка несмещенная, и ее дисперсия стремится к нулю при увеличении объема выборки, то оценка состоятельна.*

Действительно, по неравенству Чебышева,

$$P\{|\hat{\theta}_n - E\hat{\theta}_n| < \epsilon\} = P\{|\hat{\theta}_n - \theta| < \epsilon\} > 1 - \frac{D\hat{\theta}_n}{\epsilon^2} \longrightarrow 1.$$

Этому критерию удовлетворяет выборочное среднее $\bar{x}$, которое, согласно (2), является несмещенным. Покажем, что

$$\mathbf{D}\bar{x} = \frac{\sigma^2}{n}. \tag{6}$$

Поскольку элементы выборки — независимые случайные величины, то дисперсия их суммы равна сумме дисперсий, а константа $\frac{1}{n}$ выносится за знак дисперсии, будучи возведенная в квадрат,

$$\mathbf{D}\bar{x} = \mathbf{D}\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right) = \frac{1}{n^2}\sum_{k=1}^{n}\mathbf{D}x_k = \frac{1}{n^2}\sum_{k=1}^{n}\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

Таким образом, *выборочное среднее $\hat{\mu} = \bar{x}$ является несмещенной и состоятельной оценкой математического ожидания μ.* Очевидно, оценка $\hat{\mu} = x_i$ критерию состоятельности не удовлетворяет. Стандартное отклонение выборочного среднего $\frac{S}{\sqrt{n}}$ иначе называют *ошибкой среднего.*

Выясним, является ли состоятельной несмещенная оценка (5) дисперсии S^2. Для этого нужно вычислить дисперсию

$$\mathbf{D}S^2 = \frac{n^2}{(n-1)^2}\mathbf{D}m_2\,.$$

Предложение 3

$$\mathbf{D}m_2 = \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}\,. \tag{7}$$

Доказательство. Поскольку центральные моменты инвариантны относительно сдвига, поэтому, совмещая начало отсчета со средним значением, мы получим $\mu = 0$.

$$\mathbf{D}m_2 = \mathbf{E}m_2^2 - (\mathbf{E}m_2)^2\,.$$

Из (4) следует

$$\mathbf{E}m_2 = \mu_2 - \frac{\mu_2}{n}.$$

Далее $m_2^2 = a_2^2 - 2a_2\bar{x}^2 + \bar{x}^4$. Вычисляем $\mathbf{E}m_2^2$.

$$\mathbf{E}a_2^2 = \frac{1}{n^2}\mathbf{E}\left(\sum_{i=1}^n x_i^2\right)^2 = \frac{1}{n^2}\left(\sum_{i=1}^n \mathbf{E}x_i^4 + \sum_{i=1}^n\sum_{j=1,j\neq i}^n \mathbf{E}x_i^2x_j^2\right) =$$
$$= \frac{n\mu_4 + n(n-1)\mu_2^2}{n^2} = \frac{\mu_4 + (n-1)\mu_2^2}{n} = \mu_2^2 + \frac{\mu_4 - \mu_2^2}{n}\,.$$

Перейдем к основной компоненте второго слагаемого.

$$\mathbf{E}a_2\bar{x}^2 = \frac{1}{n^3}\mathbf{E}\left(\sum_{i=1}^n x_i\right)^2\sum_{i=1}^n x_i^2 = \frac{1}{n^3}\mathbf{E}\left(\sum_{i=1}^n x_i^2\sum_{i=1}^n x_i^2\right) =$$
$$= \frac{1}{n^3}\sum_{i=1}^n \mathbf{E}x_i^4 + \sum_{i=1}^n\sum_{j=1,j\neq i}^n \mathbf{E}x_i^2x_j^2 =$$

(так как слагаемые с нечетными степенями имеют нулевое математическое ожидание)

$$= \frac{n\mu_4 + n(n-1)\mu_2^2}{n^3} = \frac{\mu_4 + (n-1)\mu_2^2}{n^2} =$$
$$= \frac{\mu_2^2}{n} + \frac{\mu_4 - \mu_2^2}{n^2}.$$

Рассмотрим третье слагаемое.

$$\mathbf{E}\bar{x}^4 = \frac{1}{n^4}\mathbf{E}\left(\sum_{i=1}^{n} x_i\right)^4 =$$

$$= \frac{1}{n^4}\sum_{i1=1}^{n}\sum_{i2=1}^{n}\sum_{i3=1}^{n}\sum_{i4=1}^{n}\mathbf{E}x_{i1}x_{i2}x_{i3}x_{i4} =$$

так ненулевые слагаемые возникают только в случаях:

- $i_1 = i_2 = i_3 = i_4$;
- $i_1 = i_2, i_3 = i_4$;
- $i_1 = i_3, i_2 = i_4$;
- $i_1 = i_4, i_2 = i_3$.

$$= \frac{n\mu_4 + 3n(n-1)\mu_2^2}{n^4} = \frac{\mu_4 + 3(n-1)\mu_2^2}{n^3} = \frac{3\mu_2^2}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}.$$

Отсюда получаем

$$\mathbf{E}m_2^2 = \mu_2^2 + \frac{\mu_4 - \mu_2^2}{n} - \frac{2\mu_2^2}{n} - \frac{2\mu_4 - 2\mu_2^2}{n^2} + \frac{3\mu_2^2}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3} =$$

после приведения подобных членов

$$\mathbf{E}m_2^2 = \mu_2^2 + \frac{\mu_4 - 3\mu_2^2}{n} - \frac{2\mu_4 - 5\mu_2^2}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}.$$

Следовательно, дисперсия второго центрального момента m_2 равна

$$\mathbf{D}m_2^2 = \mathbf{E}m_2^2 - (\mathbf{E}m_2)^2 =$$

$$= \mu_2^2 + \frac{\mu_4 - 3\mu_2^2}{n} - \frac{2\mu_4 - 5\mu_2^2}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3} -$$

$$- \left(\mu_2^2 - \frac{2\mu_2^2}{n} + \frac{\mu_2^2}{n^2}\right) =$$

$$= \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3}.$$

Таким образом, дисперсия оценки m_2, а следовательно, и оценки S^2 имеет нулевой предел при увеличении объема выборки, и *оценка S^2 является состоятельной.*

2.1.3 Эффективность

Определение 4 Оценка $\hat{\theta}^1$ с меньшей дисперсией называется *эффективной*, то есть если $\mathbf{D}\hat{\theta}^1 < \mathbf{D}\hat{\theta}^2$, то $\hat{\theta}^1$ эффективнее $\hat{\theta}^2$.

Например, $\mathbf{D}\hat{\mu} = \mathbf{D}x_i = \sigma^2$, $\mathbf{D}\hat{\mu} = \mathbf{D}\bar{x} = \frac{\sigma^2}{n}$, поэтому выборочное среднее является более эффективной оценкой математического ожидания, чем произвольный элемент выборки.

Пусть $f(x,\theta)$ — генеральная плотность распределения, зависящая от параметра θ, $\mathbf{x} = (x_1, \ldots, x_n)$ — случайная выборка наблюдений, $\mathcal{L}(\mathbf{x},\theta) = f(x_1,\theta) \cdot \ldots \cdot f(x_n,\theta)$ — функция правдоподобия. Производную по параметру θ логарифма функции правдоподобия называют *информантом первого рода*:

$$s(\mathbf{x},\theta) = \frac{\partial}{\partial\theta} \ln \mathcal{L}(\mathbf{x},\theta)\,. \tag{8}$$

Свойства информанта:

- $\mathbf{E}s(\mathbf{x},\theta) = 0$.
- $\mathbf{cov}(\hat{\theta}, s(\mathbf{x},\theta)) = 1 + b'(\theta)$, где через $b(\theta) = \mathbf{E}\hat{\theta} - \theta$ обозначено смещение оценки.

$$\mathbf{E}s(\mathbf{x},\theta) = \mathbf{E}\frac{\partial}{\partial\theta}\ln\mathcal{L}(\mathbf{x},\theta) = \int\limits_{\mathcal{X}^n} \frac{\partial}{\partial\theta}\ln\mathcal{L}(\mathbf{x},\theta)\mathcal{L}(\mathbf{x},\theta)d(\mathbf{x}) =$$
$$= \int\limits_{\mathcal{X}^n} \frac{\mathcal{L}'(\mathbf{x},\theta)}{\mathcal{L}(\mathbf{x},\theta)}\mathcal{L}(\mathbf{x},\theta)d(\mathbf{x}) = \int\limits_{\mathcal{X}^n} \frac{\partial}{\partial\theta}\mathcal{L}(\mathbf{x},\theta)d(\mathbf{x}) =$$
$$= \frac{\partial}{\partial\theta}\int\limits_{\mathcal{X}^n} \mathcal{L}(\mathbf{x},\theta)d(\mathbf{x}) = 0\,.$$

Второе свойство опирается на первое утверждение.

$$\mathbf{cov}(\hat{\theta}, s(\mathbf{x},\theta)) =$$
$$= \mathbf{E}(\hat{\theta} - \mathbf{E}\hat{\theta})(s(\mathbf{x},\theta) - \mathbf{E}s(\mathbf{x},\theta)) = \mathbf{E}(\hat{\theta} - \mathbf{E}\hat{\theta})s(\mathbf{x},\theta) =$$
$$= \mathbf{E}\hat{\theta}s(\mathbf{x},\theta) - (\theta + b(\theta))\mathbf{E}s(\mathbf{x},\theta) = \mathbf{E}\hat{\theta}s(\mathbf{x},\theta)$$

$$\mathbf{E}\hat{\theta}s(\mathbf{x},\theta) = \int\limits_{\mathcal{X}^n} \hat{\theta}\frac{\partial}{\partial\theta}\ln\mathcal{L}(\mathbf{x},\theta)\mathcal{L}(\mathbf{x},\theta)d(\mathbf{x}) = \int\limits_{\mathcal{X}^n} \hat{\theta}\mathcal{L}'(\mathbf{x},\theta)d(\mathbf{x}) =$$
$$= \frac{\partial}{\partial\theta}\int\limits_{\mathcal{X}^n} \hat{\theta}\mathcal{L}(\mathbf{x},\theta)d(\mathbf{x}) = \frac{\partial}{\partial\theta}\mathbf{E}\hat{\theta} = 1 + b'(\theta)\,.$$

Предложение 4 *(неравенство Рао-Крамера)*

$$\mathbf{D}\hat{\theta} \geq \frac{(1+b'(\theta))^2}{I(\theta)},\ \textit{где}$$
$$I(\theta) = -\mathbf{E}\frac{\partial^2}{\partial\theta^2}\ln\mathcal{L}(\mathbf{x},\theta) = -\mathbf{E}\frac{\partial}{\partial\theta}s(\mathbf{x},\theta)$$

— информант второго рода.

Выражение $I(\theta)/n$ называется информационным количеством Фишера. Действительно, воспользуемся свойством корреляции.

$$\mathbf{cor}^2(\hat{\theta}, s(\mathbf{x},\theta)) = \frac{\mathbf{cov}^2(\hat{\theta}, s(\mathbf{x},\theta))}{\mathbf{D}\hat{\theta}\mathbf{D}s(\mathbf{x},\theta)} \leq 1 \iff$$
$$\iff \mathbf{D}\hat{\theta} \geq \frac{(1+b'(\theta))^2}{\mathbf{D}s(\mathbf{x},\theta)},$$
$$I(\theta) = -\mathbf{E}\frac{\partial^2}{\partial\theta^2}\ln\mathcal{L}(\mathbf{x},\theta) =$$
$$= -\int\limits_{\mathcal{X}^n}\frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta}\ln\mathcal{L}(\mathbf{x},\theta)\right)\mathcal{L}(\mathbf{x},\theta)d(\mathbf{x}) =$$
$$= -\int\limits_{\mathcal{X}^n}\frac{\mathcal{L}''\mathcal{L}-\mathcal{L}'\mathcal{L}'}{\mathcal{L}^2}\mathcal{L}d(\mathbf{x}) = \int\limits_{\mathcal{X}^n}\left(\frac{\mathcal{L}'}{\mathcal{L}}\right)^2\mathcal{L}(\mathbf{x},\theta)d(\mathbf{x}) =$$
$$= \int\limits_{\mathcal{X}^n}\left(\frac{\partial}{\partial\theta}\ln\mathcal{L}(\mathbf{x},\theta)\right)^2\mathcal{L}(\mathbf{x},\theta)d(\mathbf{x}) =$$
$$= \int\limits_{\mathcal{X}^n}\left(s(\mathbf{x},\theta)\right)^2\mathcal{L}(\mathbf{x},\theta)d(\mathbf{x}) = \mathbf{D}s(\mathbf{x},\theta)\,.$$

Определение 5 Несмещенная оценка, дисперсия которой достигает наименьшего значения, равного $I(\theta)^{-1}$, называется *эффективной. Эффективностью* несмещенной оценки, удовлетворяющей условиям регулярности, называется величина

$$e(\hat{\theta}) = \frac{1}{I(\theta)\mathbf{D}\hat{\theta}}\,. \tag{9}$$

Для эффективных оценок $e(\hat{\theta}) = 1$. *Асимптотически эффективной* называется оценка $\hat{\theta} = \hat{\theta}_n$, у которой

$$\lim_{n\longrightarrow\infty} e(\hat{\theta}_n) = 1\,.$$

ПРИМЕР 1. Экспоненциальное распределение с плотностью распределения $f(x,\alpha) = \alpha^{-1}e^{-x/\alpha}$. Математическое ожидание и дисперсия имеют вид $\mathbf{E}x_i = \alpha$, $\mathbf{D}x_i = \alpha^2$. В качестве оценки рассмотрим несмещенную оценку $\hat{\alpha} = \bar{x}$ с дисперсией $\mathbf{D}\hat{\alpha} = \frac{\alpha^2}{n}$.

$$\ln \mathcal{L}(\mathbf{x_1},\ldots,\mathbf{x_n}) = \sum_{i=1}^{n}\left(-\ln\alpha - \frac{x_i}{\alpha}\right) = -n\ln\alpha - \frac{n\bar{x}}{\alpha}.$$

Отсюда информант первого рода имеет вид

$$s(\mathbf{x},\alpha) = -\frac{n}{\alpha} + \frac{n\bar{x}}{\alpha^2}. \tag{10}$$

Дифференцируя (10) по параметру, получаем информант второго рода

$$I(\alpha) = -\mathbf{E}\left(\frac{n}{\alpha^2} - \frac{2n\bar{x}}{\alpha^3}\right) = -\frac{n}{\alpha^2} + \frac{2n}{\alpha^2} = \frac{n}{\alpha^2}\,.$$

$$I(\alpha)\mathbf{D}\hat{\alpha} = \frac{n}{\alpha^2}\cdot\frac{\alpha^2}{n} = 1\,.$$

Оценка $\hat{\alpha} = \bar{x}$ параметра α экспоненциального распределения эффективна.

ПРИМЕР 2. Нормальное распределение с логарифмом плотности

$$\ln f(x) = -\ln\sqrt{2\pi} - \frac{1}{2}\ln\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2},$$
$$\ln \mathcal{L}(x_1,\ldots,x_n) = -n\ln\sqrt{2\pi} - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Вычисляем информант первого рода

$$s(x,\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2. \tag{11}$$

Дифференцируем информант,

$$\frac{\partial}{\partial\sigma^2}s(x,\sigma^2) = \frac{n}{2\sigma^4} - \frac{2}{2\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2,$$

$$I(\sigma^2) = -\mathbf{E}\frac{\partial}{\partial\sigma^2}s(x,\sigma^2) = -\mathbf{E}\left(\frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2\right) =$$
$$= -\frac{n}{2\sigma^4} + \frac{n}{\sigma^4} = \frac{n}{2\sigma^4}, \tag{12}$$

В случае нормально распределенной генеральной совокупности, согласно (7), получаем

$$\mathbf{D}m_2 = \frac{3\sigma^4-\sigma^4}{n} - \frac{2(3\sigma^4-2\sigma^4)}{n^2} + \frac{3\sigma^4-3\sigma^4}{n^3} =$$
$$= \frac{2\sigma^4}{n} - \frac{2\sigma^4}{n^2} = \frac{2(n-1)\sigma^4}{n^2},$$
$$\mathbf{D}S^2 = \frac{n^2}{(n-1)^2}\mathbf{D}m_2 = \frac{2\sigma^4}{n-1}\,.$$

Вычисляем эффективность оценки S^2.

$$e(S^2) = \left(I(\sigma^2)\mathbf{D}S^2\right)^{-1} = \left(\frac{n}{2\sigma^4}\cdot\frac{2\sigma^4}{n-1}\right)^{-1} = \frac{n-1}{n},$$

следовательно, $\lim\limits_{n\longrightarrow\infty} e(S^2) = 1$ и S^2 является асимптотически эффективной оценкой дисперсии σ^2.

Перечислим условия регулярности.

- Множество $G = \{x : f(x,\theta) > 0\}$ не зависит от θ.
- Выражения $\int\limits_{\mathcal{X}^n} \hat{\theta}\boldsymbol{\mathcal{L}}(\mathbf{x},\theta)d(\mathbf{x})$ и $\int\limits_{\mathcal{X}^n} \boldsymbol{\mathcal{L}}(\mathbf{x},\theta)d(\mathbf{x})$ можно дифференцировать по θ под знаком интеграла.
- $I(\theta) > 0$.

2.1.4 Достаточная статистика (л.3)

Определение 6 Статистика $\hat{\theta}(\mathbf{x})$ для параметра θ является *достаточной*, если условная вероятность выборки при данном значении $\hat{\theta}$ не зависит от параметра θ.

Например, выборка (x_1, x_2) бернуллиевской случайной величины. Имеем недостаточную статистику x_1 и достаточную $x_1 + x_2$, так как

$$P\{\xi_1 = x_1, \xi_2 = x_2\} = p^{x_1+x_2} q^{2-x_1-x_2},$$
$$P\{\xi_1 = x_1, \xi_2 = x_2 | \xi_1 + \xi_2 = 1\} = \frac{pq}{2pq} = 0.5,$$
$$P\{\xi_1 = x_1, \xi_2 = x_2 | \xi_1 = 1\} = \frac{p^{1+x_2} q^{2-1-x_2}}{p} = p^{x_2} q^{1-x_2}\,.$$

Теорема 3 *(факторизации). Статистика $\hat{\theta}(\mathbf{x})$ является достаточной для параметра θ, тогда и только тогда, когда существуют такие измеримые функции h и g, что*

$$\mathcal{L}(\mathbf{x}, \theta) = h(\mathbf{x}) g\left(\hat{\theta}(\mathbf{x}), \theta\right)\,. \tag{13}$$

Доказательство. Пусть справедливо выражение (13). Тогда

$$P\{\xi_1 = x_1, \ldots, \xi_n = x_n | \hat{\theta}(\mathbf{x}) = \hat{\theta}\} =$$
$$= \frac{\mathcal{L}(\mathbf{x}, \theta)}{P\{\hat{\theta}(\mathbf{x}) = \hat{\theta}\}} = \frac{\mathcal{L}(\mathbf{x}, \theta)}{\int\limits_G \mathcal{L}(\mathbf{x}, \theta)\, d\mathbf{x}}\,,$$

где $G = \{\mathbf{x} : \hat{\theta}(\mathbf{x}) = \hat{\theta}\}$. Подставляем $\mathcal{L}(\mathbf{x}, \theta) = h(\mathbf{x}) g\left(\hat{\theta}(\mathbf{x}), \theta\right)$.

$$\frac{h(\mathbf{x}) g\left(\hat{\theta}(\mathbf{x}), \theta\right)}{\int\limits_G h(\mathbf{x}) g\left(\hat{\theta}(\mathbf{x}), \theta\right) d\mathbf{x}} = \frac{h(\mathbf{x})}{\int\limits_G h(\mathbf{x}) d\mathbf{x}}\,.$$

Последнее выражение от θ не зависит, следовательно, $\hat{\theta}$ является достаточной. В обратную сторону, при достаточной статистике первый сомножитель в выражении

$$\mathcal{L}(\mathbf{x}, \theta) = P\{\xi_1 = x_1, \ldots, \xi_n = x_n | \hat{\theta}(\mathbf{x}) = \hat{\theta}\} P\{\hat{\theta}(\mathbf{x}) = \hat{\theta}\}$$

от θ не зависит, и может рассматриваться как $h(\mathbf{x})$, а второй можно взять как функцию $g\left(\hat{\theta}(\mathbf{x}), \theta\right)$.

ПРИМЕР 1. Пусть $\mathbf{x} = (x_1, \ldots, x_n)$ — выборка наблюдений за бернуллиевской случайной величиной с вероятностью успеха p. Функция правдоподобия имеет вид

$$\mathcal{L}(\mathbf{x}, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} =$$
$$= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} = p^{n\bar{x}}(1-p)^{n-n\bar{x}}\,.$$

Статистика $\hat{p} = \bar{x}$ является достаточной, так как $h(\mathbf{x}) = 1$, $g\left(\bar{x}, p\right) = p^{n\bar{x}}(1-p)^{n-n\bar{x}}$.

ПРИМЕР 2. Функция правдоподобия распределения Пуассона имеет вид

$$\mathcal{L}(\mathbf{x}, \lambda) = \prod_{i=1}^{n}\left(\frac{\lambda^{x_i}}{x_i!}e^{-\lambda}\right) = e^{-n\lambda}\lambda^{\sum\limits_{i=1}^{n} x_i}\frac{1}{x_1!\ldots x_n!}\,.$$

Достаточной статистикой оказывается $T = \sum\limits_{i=1}^{n} x_i$, так как

$$h(\mathbf{x}) = \frac{1}{x_1!\ldots x_n!}, \quad g(T, \lambda) = e^{-n\lambda}\lambda^{T}\,.$$

ПРИМЕР 3. Функция правдоподобия экспоненциального распределения имеет вид

$$\mathcal{L}(\mathbf{x}, \alpha) = \prod_{i=1}^{n}\left(\frac{1}{\alpha}e^{-\frac{x_i}{\alpha}}\right) = \frac{1}{\alpha^n}e^{-\sum\limits_{i=1}^{n}\frac{x_i}{\alpha}} = \frac{1}{\alpha^n}e^{-\frac{n\bar{x}}{\alpha}}\,,$$

статистика $\hat{\alpha} = \bar{x}$ является достаточной, так как $h(\mathbf{x}) = 1$, $g(\bar{x}, \alpha) = \alpha^{-n}e^{-\frac{n\bar{x}}{\alpha}}$. Заметим, что статистика $\hat{\alpha} = \bar{x}_1$ достаточной не является, так как если взять в качестве $g(x_1, \alpha) = \alpha^{-n}e^{-\frac{x_1}{\alpha}}$, то оставшийся сомножитель $h(\mathbf{x}) = e^{-\sum\limits_{i=2}^{n} x_i/\alpha}$ будет зависеть от α.

Достаточная статистика $\hat{\theta}_*$ называется *минимально достаточной*, если для каждой достаточной статистики $\hat{\theta}$ существует неслучайная измеряемая функция g, что почти всюду $\hat{\theta}_* = g(\hat{\theta})$.

2.2 Метод моментов

Рассматривая количество моментов, равное количеству оцениваемых параметров, и приравнивая выборочные моменты к соответствующим моментам распределения, являющимся функциями от параметров, можно получить искомые оценки.

ПРИМЕР 1. Пусть имеется $x_1, \ldots, x_N$ выборка наблюдений биномиальной случайной величины с параметрами n и p. $\alpha_1 = np, \quad \mu_2 = npq$. Приравнивая эти выражения к соответствующим выборочным моментам $\bar{x} = \frac{1}{N}\sum\limits_{i=1}^{N} x_i$ и $m_2 = \frac{1}{N}\sum\limits_{i=1}^{N}(x_i - \bar{x})^2$, получаем

$$\hat{q} = \frac{m_2}{\bar{x}}, \quad \hat{n} = \frac{\bar{x}}{1-\hat{q}} = \frac{\bar{x}^2}{\bar{x} - m_2}\,.$$

ПРИМЕР 2. Рассмотрим случайную величину, имеющую сложно пуассоновское распределение, а именно $\xi = \eta_1 + \ldots + \eta_\tau$, где случайная величина τ имеет пуассоновское распределение с параметром λ, а величины $\eta_1, \ldots, \eta_\tau$ независимые, одинаково распределенные, например, биномиальные с вероятностью успеха p и числом независимых испытаний n.

Производящая функция дискретного распределения

$$\xi : \begin{pmatrix} 0 & 1 & \ldots & n & \ldots \\ p_0 & p_1 & \ldots & p_n & \ldots \end{pmatrix}$$

определяется в виде степенного ряда $f(s) = \sum\limits_{j=0}^{\infty} p_j s^j$. Для пуассоновской и биномиальной величин имеем соответственно

$$f(s) = \sum_{j=0}^{\infty} \frac{\lambda^j s^j}{j!} e^{-\lambda} = e^{-\lambda+\lambda s}\,,$$

$$g(s) = \sum_{k=0}^{n} C_n^k p^k q^{n-k} s^k = (q+ps)^n\,.$$

Производящая функция сложного распределения имеет вид суперпозиции

$$h(s) = f(g(s)) = e^{-\lambda+\lambda(q+ps)^n}\,.$$

Воспользуемся свойствами производящих функций:

$$h'(s) = \sum_{j=0}^{\infty} j p_j s^{j-1}, \quad h'(1) = \mathbf{E}\xi = \alpha_1, \quad h''(s) = \sum_{j=0}^{\infty} j(j-1) p_j s^{j-2},$$

$$h''(1) = \mathbf{E}\xi^2 - \mathbf{E}\xi = \alpha_2 - \alpha_1,$$

$$\alpha_1 = h'(1), \quad \alpha_2 = h''(1) + h'(1), \quad \mu_2 = h''(1) + h'(1) - (h'(1))^2\,.$$

Пусть n известно. Тогда для получения оценок по методу моментов достаточно первых двух моментов.

$$h'(s) = e^{-\lambda+\lambda(q+ps)^n} \cdot \lambda n(q+ps)^{n-1}p,$$
$$h''(s) = e^{-\lambda+\lambda(q+ps)^n} \cdot (\lambda n(q+ps)^{n-1}p)^2 +$$
$$+e^{-\lambda+\lambda(q+ps)^n} \cdot \lambda np(n-1)(q+ps)^{n-2}p,$$
$$h'(1) = \lambda np = \alpha_1\,,$$
$$h''(1) = (\lambda np)^2 + \lambda np^2(n-1),$$
$$\mu_2 = h''(1) + h'(1) - (h'(1))^2 =$$
$$= (\lambda np)^2 + \lambda n^2p^2 - \lambda np^2 + \lambda np - (\lambda np)^2 = \lambda n^2p^2 + \lambda npq;$$
$$\epsilon = \frac{\mu_2}{\alpha_1} = np + 1 - p = p(n-1) + 1;$$
$$p = \frac{\epsilon - 1}{n-1}, \quad \lambda = \frac{\alpha_1}{np}\,.$$

Подставляя выборочные оценки $\bar{x}$ и m_2, получаем оценки $\hat{p}, \hat{\lambda}$. Самостоятельно предлагается найти выражения μ_3 и μ_4 для этого распределения.

Смещение оценок по методу моментов устраняется введением некоторых поправок. Для проверки состоятельности необходимы значения дисперсии этих оценок.

Теорема 4 *Пусть в некоторой окрестности точки $m_\nu = \mu_\nu$ и $m_\varrho = \mu_\varrho$ функция $H(m_\nu, m_\varrho)$ непрерывна м имеет первые и вторые производные по аргументам m_ν, m_ϱ, а также для всевозможных x_i выполнено неравенство $|H| < Cn^p$, где C и p — некоторые постоянные. Тогда если обозначить через H_0, H_1, H_2 значения функции $H(m_\nu, m_\varrho)$ и ее первых частных производных в точке $m_\nu = \mu_\nu$ и $m_\varrho = \mu_\varrho$, то среднее и дисперсия случайной величины $H(m_\nu, m_\varrho)$ будут определяться формулами*

$$\mathbf{E}H = H_0 + O\left(\frac{1}{n}\right),$$
$$\mathbf{D}H = \mu_2(m_\nu)H_1^2 + 2\mu_{11}(m_\nu, m_\varrho)H_1H_2 + \tag{14}$$
$$+\mu_2(m_\varrho)H_2^2 + O\left(\frac{1}{n^{\frac{3}{2}}}\right).$$

Применим эту теорему для вычисления дисперсии рассеяния

$$e = H(\bar{x}, m_2) = \frac{m_2}{\bar{x}}.$$

$$H_1 = \frac{\partial}{\partial \bar{x}} H(\bar{x}, m_2) = -\frac{m_2}{\bar{x}^2} \mid_{m_2=\mu_2, \bar{x}=\mu} = -\frac{\mu_2}{\mu^2},$$
$$H_2 = \frac{\partial}{\partial m_2} H(\bar{x}, m_2) = \frac{1}{\bar{x}} \mid_{\bar{x}=\mu} = \frac{1}{\mu}.$$

Из (6) и (7)

$$\mu_2(\bar{x}) = \mathbf{D}\bar{x} = \frac{\mu_2}{n}, \quad \mu_2(m_2) = \frac{\mu_4 - \mu_2^2}{n} + O\left(\frac{1}{n^2}\right).$$

Вычислим смешанный центральный момент

$$\mu_{11}(\bar{x}, m_2) = \mathbf{E}\left((\bar{x} - \mu)\left(m_2 - \frac{n-1}{n}\mu_2\right)\right) = \mathbf{E}\left((\bar{x} - \mu)m_2\right).$$

Последнее выражение не зависит от положения начала отсчета, поэтому положим $\mu = 0$.

$$\mu_{11}(\bar{x}, m_2) = \mathbf{E}(\bar{x}m_2) = \mathbf{E}(\bar{x}a_2) - \mathbf{E}(\bar{x}^3) =$$
$$= \frac{\mu_3}{n} - \frac{\mu_3}{n^2} = \frac{n-1}{n^2}\mu_3.$$

Таким образом, с точностью до $O\left(\frac{1}{n^{\frac{3}{2}}}\right)$

$$\mathbf{D}e = \frac{\mu_2}{n}\left(-\frac{\mu_2}{\mu^2}\right)^2 + 2 \cdot \frac{n-1}{n^2}\mu_3\left(-\frac{\mu_2}{\mu^2}\right) \cdot \frac{1}{\mu} +$$
$$+\frac{\mu_4 - \mu_2^2}{n} \cdot \frac{1}{\mu^2} = \frac{\mu_2^3}{n\mu^4} - \frac{2\mu_3\mu_2}{n\mu^3} + \frac{\mu_4 - \mu_2^2}{n\mu^2}.$$

Например, для нормального распределения приближенное выражение для дисперсии рассеяния принимает вид

$$\mathbf{D}e = \frac{\sigma^6}{n\mu^4} + \frac{2\sigma^4}{n\mu^2}.$$

2.3 Метод максимального правдоподобия (л.4)

2.3.1 Асимптотические эффективность и нормальность

Если выборочные значения фиксированы, то функция правдоподобия $\mathcal{L}(x_1, \ldots, x_n, \theta) = \mathcal{L}(\mathbf{x}, \theta)$ является функцией от единственного аргумента θ. Метод максимального правдоподобия заключается в том, что в качестве оценки параметра принимается значение $\hat{\theta}$, при котором функция правдоподобия достигает своего максимального значения. Поскольку $\ln \mathcal{L}(\mathbf{x}, \theta)$ достигает максимума в той же точке, что и $\mathcal{L}(\mathbf{x}, \theta)$, следует решать относительно θ *уравнение правдоподобия.*

$$\frac{\partial \ln \mathcal{L}(\mathbf{x}, \theta)}{\partial \theta} = 0\,. \tag{15}$$

Каждое решение уравнения (15) будем называть *оценкой максимального правдоподобия.*

Предложение 5 *При некоторых общих условиях уравнение правдоподобия имеет решение, сходящееся по вероятности к истинному значению параметра при* $n \to \infty$. *Оно является асимптотически нормальной и асимптотически эффективной оценкой для параметра* θ.

Предположим, что A — некоторый невырожденный интервал, и выполнены условия:

1. При каждом $\theta \in A$ для почти всех x существуют производные $\frac{\partial \ln f}{\partial \theta}$, $\frac{\partial^2 \ln f}{\partial \theta^2}$, $\frac{\partial^3 \ln f}{\partial \theta^3}$.

2. При каждом $\theta \in A$ имеем $\left|\frac{\partial f}{\partial \theta}\right| < F_1(x)$, $\left|\frac{\partial^2 f}{\partial \theta^2}\right| < F_2(x)$, $\left|\frac{\partial^3 \ln f}{\partial \theta^3}\right| < H(x)$, $F_1(x)$ и $F_2(x)$ интегрируемы на $(-\infty; +\infty)$, $\int\limits_{-\infty}^{+\infty} H(x) f(x, \theta) dx < M$, причем M не зависит от θ.

3. При каждом $\theta \in A$ интеграл $\int\limits_{-\infty}^{+\infty} \left(\frac{\partial \ln f}{\partial \theta}\right)^2 f(x, \theta) dx$ конечен и положителен.

Пусть θ_0 — истинное значение параметра и θ_0 — внутренняя точка интервала A. Покажем, что уравнение правдоподобия имеет решение, сходящееся по вероятности к θ_0.

$$\frac{\partial \ln f}{\partial \theta} = \left(\frac{\partial \ln f}{\partial \theta}\right)_0 + (\theta - \theta_0)\left(\frac{\partial^2 \ln f}{\partial \theta^2}\right)_0 + \frac{1}{2}\Theta(\theta - \theta_0)^2 H(x),$$

где $|\Theta| < 1$, а индексом 0 указано, что для каждого $\theta \in A$ следует положить $\theta = \theta_0$. Умножим уравнение (15) на $\frac{1}{n}$.

$$\frac{1}{n}\frac{\partial \ln \mathcal{L}}{\partial \theta} = B_0 + B_1(\theta - \theta_0) + \frac{1}{2}\Theta B_2(\theta - \theta_0)^2 = 0, \qquad (16)$$

где, обозначая через $f_i = f(x_i, \theta)$, имеем

$$B_0 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_i}{\partial \theta}\right)_0, \quad B_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial^2 \ln f_i}{\partial \theta^2}\right)_0,$$
$$B_2 = \frac{1}{n}\sum_{i=1}^{n} H(x_i)\,.$$

Заметим, что по аналогии со свойствами информанта $s(\mathbf{x}, \theta)$

$$\mathbf{E}\left(\frac{\partial \ln f_i}{\partial \theta}\right)_0 = \int_{-\infty}^{+\infty}\left(\frac{\partial \ln f_i}{\partial \theta}\right)_0 f(x_i, \theta_0)dx_i = 0\,,$$

$$\mathbf{E}\left(\frac{\partial^2 \ln f}{\partial \theta^2}\right)_0 = \int_{-\infty}^{+\infty}\left(\frac{1}{f}\frac{\partial^2 f}{\partial \theta^2} - \left(\frac{1}{f}\frac{\partial f}{\partial \theta}\right)^2\right)_0 f(x_i, \theta_0)dx =$$
$$= -\mathbf{E}\left(\frac{\partial \ln f}{\partial \theta}\right)_0^2 = -k^2\,.$$

Итак, величина B_0 есть среднее арифметическое n независимых одинаково распределенных случайных величин с нулевым математическим ожиданием. Из теоремы Хинчина следует, что B_0 сходится по вероятности к нулю. Аналогично B_1 сходится по вероятности к $-k^2$, а B_2 сходится к неотрицательному значению $\mathbf{E}H(x) < M$.

Покажем, что для произвольно малых δ и ϵ уравнение правдоподобия имеет с вероятностью, превышающей $1-\epsilon$, корень, заключенный в пределах $\theta_0 \pm \delta$, если $n > n(\delta, \epsilon)$. Пусть S множество точек $(x_1, \ldots, x_n)$, для которых выполнены неравенства:

$$|B_0| < \delta^2, \quad B_1 < -\frac{k^2}{2}, \quad |B_2| < 2M\,.$$

$$\forall\ n > n(\delta,\epsilon) \quad P_0 = P\left\{|B_0| > \delta^2\right\} < \frac{\epsilon}{3},$$

$$P_1 = P\left\{B_1 > -\frac{k^2}{2}\right\} < \frac{\epsilon}{3}, \quad P_2 = P\left\{|B_2| > 2M\right\} < \frac{\epsilon}{3}\,.$$

Дополнительное множество $\bar{S}$ состоит из точек, для которых справедливо хотя бы одно из этих неравенств, поэтому

$$P(\bar{S}) < P_0 + P_1 + P_2 < \epsilon,$$

$$P(S) > 1 - \epsilon\,.$$

При $\theta = \theta_0 \pm \delta$ правая часть уравнения (15) имеет вид

$$B_0 \pm B_1\delta + \frac{1}{2}\Theta B_2\delta^2\,.$$

В каждой точке $\mathbf{x} \in S$

$$|B_0 + \frac{1}{2}\Theta B_2\delta^2| < \delta^2 + \frac{1}{2}\cdot 2M\delta^2 = (M+1)\delta^2,$$

$$B_1\delta < -\frac{k^2\delta}{2}\,.$$

При $\theta = \theta_0 - \delta$ и при $\delta^2(M+1) < \frac{k^2\delta}{2}$ имеем

$$\frac{1}{n}\frac{\partial \ln \mathcal{L}}{\partial \theta} = \left(B_0 + \frac{1}{2}\Theta B_2\delta^2\right) + (-B_1\delta) > 0,$$

так как $(-B_1\delta) > \dfrac{k^2\delta}{2} > \delta^2(M+1)$, а при $\theta = \theta_0 + \delta$

$$\frac{1}{n}\frac{\partial \ln \mathcal{L}}{\partial \theta} = \left(B_0 + \frac{1}{2}\Theta B_2\delta^2\right) + B_1\delta < 0,$$

так как $B_1\delta < -\dfrac{k^2\delta}{2} < -\delta^2(M+1) < -\left(B_0 + \dfrac{1}{2}\Theta B_2\delta^2\right)$.

Итак, существование решения θ^* уравнения правдоподобия установлено. Из уравнения (16)

$$B_0 + B_1(\theta - \theta_0) + \frac{1}{2}\Theta B_2(\theta - \theta_0)^2 = 0$$

выразим

$$B_0 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f_i}{\partial \theta}\right)_0 = (\theta^* - \theta)\left(-B_1 - \frac{1}{2}\Theta B_2(\theta^* - \theta)\right),$$

$$k\sqrt{n}(\theta^* - \theta_0) = \frac{\frac{1}{k\sqrt{n}}\sum\limits_{i=1}^{n}\left(\frac{\partial \ln f_i}{\partial \theta}\right)_0}{-\frac{B_1}{k^2} - \frac{1}{2}\Theta B_2\frac{\theta^* - \theta_0}{k^2}}.$$

Знаменатель дроби сходится по вероятности к 1. Так как

$$\left(\frac{\partial \ln f_i}{\partial \theta}\right)_0$$

имеет математическое ожидание 0 и стандартное отклонение k, сумма $\sum\limits_{i=1}^{n}\left(\frac{\partial \ln f_i}{\partial \theta}\right)_0$ асимптотически нормальна $\mathcal{N}(0, k\sqrt{n})$.

Следовательно, числитель дроби асимптотически нормален $\mathcal{N}(0,1)$, а оценка θ^* асимптотически нормальна $\mathcal{N}(0, \frac{1}{k\sqrt{n}})$, где

$$k^2 = \mathbf{E}\left(\frac{\partial \ln f}{\partial \theta}\right)_0^2.$$

Асимптотическая эффективность равна

$$e(\theta^*) = \left(\frac{1}{nk^2}\cdot I(\theta)\right)^{-1} = 1,$$

$$\text{так как } \; I(\theta) = -\mathbf{E}\frac{\partial^2}{\partial \theta^2}\ln \mathcal{L}(\mathbf{x}, \theta) = nk^2.$$

2.3.2 ОМП параметров нормального распределения

Пусть имеется выборка $x_1, \ldots, x_n$ с генеральной плотностью $\mathcal{N}(\mu, \sigma)$. Функция правдоподобия имеет вид

$$\mathcal{L}(x_1, \ldots, x_n) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}},$$

$$\ln \mathcal{L}(x_1, \ldots, x_n) = -\frac{n}{2}\left(\ln(2\pi) + \ln \sigma^2\right) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\,.$$

В результате дифференцирования по параметрам получаем систему нормальных уравнений.

$$\begin{cases} \frac{\partial \ln \mathcal{L}(\mathbf{x})}{\partial \mu} = -\sum\limits_{i=1}^{n} \frac{2(x_i-\mu)\cdot(-1)}{2\sigma^2} = \sum\limits_{i=1}^{n} \frac{(x_i-\mu)}{\sigma^2} = 0 \\ \frac{\partial \ln \mathcal{L}(\mathbf{x})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \sum\limits_{i=1}^{n} \frac{(x_i-\mu)^2}{2(\sigma^2)^2} = 0 \end{cases}$$

Решением этой системы уравнений являются оценки вида

$$\begin{cases} \hat{\mu} = \bar{x}, \\ \hat{\sigma}^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \hat{\mu})^2 = m_2. \end{cases}$$

Оценка $\hat{\sigma}^2 = m_2$ асимптотически эффективна.

$$\frac{\partial^2 \ln \mathcal{L}(\mathbf{x})}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{\sum\limits_{i=1}^{n}(x_i - \mu)^2}{\sigma^6},$$

$$I(\sigma^2) = -\mathbf{E}\frac{\partial^2 \ln \mathcal{L}(\mathbf{x})}{\partial (\sigma^2)^2} = -\frac{n}{2\sigma^4} + \frac{n\sigma^2}{2\sigma^6} = \frac{n}{2\sigma^4},$$

$$\mathbf{D}m_2 = \frac{\mu_4 - \mu_2^2}{n} + O\left(\frac{1}{n^2}\right) = \frac{2\sigma^4}{n} + O\left(\frac{1}{n^2}\right),$$

$$e(m_2) = \left(I(\sigma^2)\mathbf{D}m_2\right)^{-1} =$$

$$= \left(\frac{n}{2\sigma^4}\left(\frac{2\sigma^4}{n} + O\left(\frac{1}{n^2}\right)\right)\right)^{-1} = \left(1 + O\left(\frac{1}{n}\right)\right)^{-1}_{n\to\infty} \to 1\,.$$

Сравнивая одинаково асимптотически эффективные оценки S^2 и m_2, заметим, что S^2 обладает несмещенностью, а m_2 имеет меньшую дисперсию, то есть более эффективна.

3 Теорема Пирсона и ее приложения

3.1 χ^2-квадрат распределение (л.5)

3.1.1 Гамма-функция

Гамма-функцией называется несобственный интеграл

$$\Gamma(\lambda) = \int_0^\infty x^{\lambda-1} e^{-x} dx, \quad \lambda > 0\,.$$

Свойства гамма-функции.

1. $\Gamma(\lambda+1) = \lambda\Gamma(\lambda)$;
2. $\Gamma(n+1) = n!, \;\; n \in \mathbf{N}$;
3. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

Первое утверждение следует из интегрирования по частям.

$$\Gamma(\lambda+1) = \int_0^\infty \underbrace{x^\lambda}_{u} \underbrace{e^{-x}}_{v'} dx = uv\,|_0^\infty - \int_0^\infty u'v dx =$$

$$= x^\lambda \left(-e^{-x}\right)|_0^\infty - \int_0^\infty \lambda x^{\lambda-1}\left(-e^{-x}\right) dx = \lambda\Gamma(\lambda)\,,$$

$$\Gamma(1) = \int_0^\infty x^0 e^{-x} dx = -e^{-x}\,|_0^\infty = 1\,.$$

Если $\lambda \in \mathbf{N}$, то повторное применение $\Gamma(n+1) = n\Gamma(n)$ приведет к $\Gamma(n+1) = n!$. Последнее получаем из интеграла Пуассона.

$$\int_0^\infty \int_0^\infty e^{-x^2-y^2} dxdy = \left(\int_0^\infty e^{-x^2} dx\right)^2 = \int_0^{\frac{\pi}{2}} d\phi \int_0^\infty e^{-r^2} r dr = \frac{\pi}{4}\,.$$

Используя замену $\sqrt{x} = t, \;\; x = t^2, \;\; dx = 2tdt$, получаем

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty \frac{1}{\sqrt{x}} e^{-x} dx = \int_0^\infty \frac{1}{t} e^{-t^2} 2tdt = 2\frac{\sqrt{\pi}}{2} = \sqrt{\pi}\,.$$

3.1.2 Гамма-распределение

Нетрудно показать, что

$$\int\limits_0^\infty x^{\lambda-1}e^{-\alpha x}dx = \int\limits_0^\infty \frac{(\alpha x)^{\lambda-1}}{\alpha^{\lambda-1}}e^{-\alpha x}d(\alpha x)\frac{1}{\alpha} = \frac{\Gamma(\lambda)}{\alpha^\lambda},$$

отсюда получаем, что функция вида

$$\gamma(x,\alpha,\lambda) = \begin{cases} \frac{\alpha^\lambda}{\Gamma(\lambda)}x^{\lambda-1}e^{-\alpha x}, & x>0, \\ 0, & x\le 0, \end{cases}$$

обладает свойством плотности распределения случайной величины ξ с характеристической функцией (х.ф.)

$$\psi(t) = \mathbf{E}e^{it\xi} = \int\limits_{-\infty}^{\infty} e^{itx}\gamma(x,\alpha,\lambda) =$$

$$= \frac{\alpha^\lambda}{\Gamma(\lambda)}\int\limits_0^\infty x^{\lambda-1}e^{-(\alpha-it)x}dx = \frac{\alpha^\lambda}{\Gamma(\lambda)}\cdot\frac{\Gamma(\lambda)}{(\alpha-it)^\lambda} = \left(1-\frac{it}{\alpha}\right)^{-\lambda}.$$

Параметры α и λ называются соответственно параметрами *масштаба* и *формы.* Из того, что х.ф. суммы независимых случайных величин равна произведению х.ф., получаем, что если случайные величины $\xi_1 \sim \gamma(x,\alpha,\lambda_1)$ и $\xi_2 \sim \gamma(x,\alpha,\lambda_2)$ независимы, то $\xi_1+\xi_2 \sim \gamma(x,\alpha,\lambda_1+\lambda_2)$. Для вычисления моментов воспользуемся свойством характеристической функции $\psi^{(\nu)}(0) = i^\nu\alpha_\nu$.

$$\psi'(t) = -\lambda\left(1-\frac{it}{\alpha}\right)^{-\lambda-1}\left(-\frac{i}{\alpha}\right),$$

$$\psi''(t) = (-\lambda)(-\lambda-1)\left(1-\frac{it}{\alpha}\right)^{-\lambda-2}\left(-\frac{i}{\alpha}\right)^2,$$

$$\alpha_1 = \frac{\psi'(0)}{i} = \frac{\lambda}{\alpha},$$

$$\alpha_2 = \frac{\psi''(0)}{i^2} = \frac{1}{i^2}\lambda(\lambda+1)\left(-\frac{i}{\alpha}\right)^2 = \frac{\lambda^2+\lambda}{\alpha^2},$$

$$\mu_2 = \alpha_2 - \alpha_1^2 = \frac{\lambda}{\alpha^2}.$$

3.1.3 Распределение χ^2 как частный случай γ

Существенным частным случаем гамма-распределения является распределение *хи-квадрат*. Говорят, что случайная величина η имеет распределение $\chi^2(n)$ с n степенями свободы, если она распределена также, как сумма квадратов n независимых стандартно нормально распределенных случайных величин.

Пусть $\xi \sim \mathcal{N}(0,1)$. Вычислим функцию $F_\eta(x)$ и плотность $f_\eta(x)$ распределения случайной величины $\eta = \xi^2$.

$$F_\eta(x) = P\{\eta \le x\} = P\{\xi^2 \le x\} = P\{-\sqrt{x} < \xi \le \sqrt{x}\} = \\ = \begin{cases} F_\xi(\sqrt{x}) - F_\xi(-\sqrt{x}), & x > 0, \\ 0, & x \le 0. \end{cases}$$

$$f_\eta(x) = F'_\eta(x) = \begin{cases} \frac{F'_\xi(\sqrt{x}) + F'_\xi(-\sqrt{x})}{2\sqrt{x}}, & x > 0. \\ 0, & x \le 0. \end{cases}$$

Так как для $\mathcal{N}(0,1)$ закона распределения имеет место $f_\xi(x) = F'_\xi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_2}{2}}$, для $x > 0$ плотность квадрата стандартно нормально распределенной величины есть $\gamma\left(x, \frac{1}{2}, \frac{1}{2}\right)$ — плотность гамма распределения

$$f_\eta(x) = \frac{1}{2\sqrt{x}} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{x}{2}} + \frac{1}{\sqrt{2\pi}} e^{-\frac{x}{2}} \right) = \\ = \frac{1}{\sqrt{2\pi x}} e^{-\frac{x}{2}} = \frac{\left(\frac{1}{2}\right)^{\frac{1}{2}}}{\Gamma\left(\frac{1}{2}\right)} x^{-\frac{1}{2}} e^{-\frac{x}{2}}$$

с х.ф. $\psi(t) = (1 - 2it)^{-\frac{1}{2}}$. Следовательно, если речь идет о сумме n независимых величин ξ^2, то получаем характеристическую функцию $\psi_n(t) = (1 - 2it)^{-\frac{n}{2}}$ гамма распределенной величины с параметром масштаба $\alpha = \frac{1}{2}$ и формы $\lambda = \frac{n}{2}$. Моменты получают достаточно простое выражение:

$$\alpha_1 = \frac{\lambda}{\alpha} = \frac{\frac{n}{2}}{\frac{1}{2}} = n, \quad \mu_2 = \frac{\lambda}{\alpha^2} = \frac{\frac{n}{2}}{\frac{1}{2^2}} = 2n\,.$$

3.2 Теорема о согласии распределений

3.2.1 Случай известных параметров

Нулевая гипотеза H_0 заключается в том, что выборка наблюдений $x_1, \ldots, x_n$ относится к случайной величине ξ с генеральной функцией распределения $F(x)$. Множество S значений случайной величины разбито на r непересекающихся интервалов $S_1, \ldots, S_r$.

$$\cup_{i=1}^{r} S_i = S, \quad S_i \cap S_j = \emptyset, \ i \neq j,$$
$$p_i = P\{\xi \in S_i\} > 0, \quad \sum_{i=1}^{r} p_i = 1.$$

Соответствующие эмпирические частоты равны ν_i, $\sum\limits_{i=1}^{r} \nu_i = n$. В качестве меры расхождения между эмпирическим и генеральным распределением рассматривается величина

$$\sum_{i=1}^{r} c_i \left(\frac{\nu_i}{n} - p_i \right)^2,$$

где, согласно Пирсону, коэффициенты имеют вид $c_i = \frac{n}{p_i}$.

Теорема 5 *При справедливости H_0 статистика*

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i} \tag{17}$$

имеет распределение $\chi^2(r-1)$ хи-квадрат с числом степеней свободы, равным $r-1$.

Мультиномиальное распределение является обобщением биномиального, в котором вместо двух исходов (успех или неудача) рассматривается r вариантов [1]. Пусть $\nu_1, \ldots, \nu_r$ - случайные величины, означающие количества разных видов исходов,

[1] Например, имеются три исхода: „да“, „нет“, „не знаю“, которые осуществляются с вероятностями p_1, p_2, p_3, и n независимых испытаний. Количество вариантов получить n_1 ответов „да“, n_2 ответов „нет“ и n_3 ответов „не знаю“, где $n_1 + n_2 + n_3 = n$, равно $C_n^{n_1} C_{n-n_1}^{n_2} C_{n-n_1-n_2}^{n_3} = \frac{n!}{n_1! n_2! n_3!}$.

$\nu_1 + \ldots + \nu_r = n$. Вероятность i-го исхода равна p_i, $\sum\limits_{i=1}^{r} p_i = 1$.

$$P\{\nu_1 = n_1, \ldots, \nu_r = n_r\} = \frac{n!}{n_1! \ldots n_r!} p_1^{n_1} \ldots p_r^{n_r} .$$

Характеристическая функция мультиномиального распределения имеет вид:

$$\Phi_\nu(t_1, \ldots, t_r) = \mathbf{E} e^{it^T \nu} = \mathbf{E} e^{i(t_1\nu_1 + \ldots + t_r\nu_r)} = \\ = (p_1 e^{it_1} + \ldots + p_r e^{it_r})^n .$$

В частности, $\phi(t_1) = \Phi_\nu(t_1, 0, \ldots, 0) =$

$$= (p_1 e^{it_1} + p_2 e^{i\cdot 0} + \ldots + p_r e^{i\cdot 0})^n = (p_1 e^{it_1} + p_2 + \ldots + p_r)^n = \\ = (p_1 e^{it_1} + (1 - p_1))^n$$

— характеристическая функция биномиального распределения.

Обозначим через $x_i = \frac{\nu_i - np_i}{\sqrt{np_i}}$ и покажем, что $\sum\limits_{i=1}^{r} x_i \sqrt{np_i} = 0$. Действительно, $\nu_i = np_i + x_i\sqrt{np_i}$,

$$n = \sum_{i=1}^{r} \nu_i = \sum_{i=1}^{r} (np_i + x_i\sqrt{np_i}) = n + \sum_{i=1}^{r} x_i \sqrt{np_i} .$$

Рассмотрим характеристическую функцию совместного распределения величин $x_1, \ldots, x_r$

$$\Phi_x(t_1, \ldots, t_r) = \mathbf{E} e^{i(t_1 x_1 + \ldots + t_r x_r)} = \\ = \mathbf{E} e^{i(t_1 \frac{\nu_1}{\sqrt{np_1}} + \ldots + t_r \frac{\nu_1 r}{\sqrt{np_r}}) - i(t_1\sqrt{np_1} + \ldots + t_r\sqrt{np_r})} = \\ = e^{-i\sqrt{n}\sum_k t_k\sqrt{p_k}} \Phi_\nu \left(\frac{t_1}{\sqrt{np_1}}, \ldots \frac{t_r}{\sqrt{np_r}} \right) = \\ = e^{-i\sqrt{n}\sum_k t_k\sqrt{p_k}} \left(p_1 e^{\frac{it_1}{\sqrt{np_1}}} + \ldots + p_r e^{\frac{it_r}{\sqrt{np_r}}} \right)^n .$$

Прологарифмируем это выражение,

$$\ln \Phi_x(t) = -i\sqrt{n} \sum_k t_k \sqrt{p_k} + n \ln(p_1 e^{\frac{it_1}{\sqrt{np_1}}} + \ldots + p_r e^{\frac{it_r}{\sqrt{np_r}}}),$$

и для второго слагаемого воспользуемся разложениями Тейлора

$$e^x = 1 + x + \frac{x^2}{2!} + O(x^3),\ \ln(1+x) = x - \frac{x^2}{2} + O(x^3)\,.$$

$$n\ln(p_1 e^{\frac{it_1}{\sqrt{np_1}}} + \ldots + p_r e^{\frac{it_r}{\sqrt{np_r}}}) =$$

$$= n\ln\left(\underbrace{\sum_k p_k}_{1} + \underbrace{\sum_k p_k \frac{it_k}{\sqrt{np_k}} + \frac{1}{2}\sum_k p_k \left(\frac{it_k}{\sqrt{np_k}}\right)^2 + O(n^{-\frac{3}{2}})}_{X}\right) =$$

$$= n\left(X - \frac{1}{2}X^2 + O(X^3)\right) =$$

$$= n\left(\frac{i}{\sqrt{n}}\sum_k \sqrt{p_k}t_k + \frac{1}{2n}\sum_k (it_k)^2 + O(n^{-\frac{3}{2}})\right) -$$

$$-\frac{n}{2}\left(\frac{i^2}{n}\left(\sum_k \sqrt{p_k}t_k\right)^2 + O(n^{-\frac{3}{2}})\right) + O(n^{-\frac{1}{2}})$$

При переходе к $\ln \Phi_x(t)$ первое слагаемое исчезает.

$$\ln \Phi_x(t) = -\frac{1}{2}\sum_k t_k^2 + \frac{1}{2}\left(\sum_k \sqrt{p_k}t_k\right)^2 + O(n^{-\frac{1}{2}}) =$$

$$= -\frac{1}{2}\left(\sum_k t_k^2 - \left(\sum_k \sqrt{p_k}t_k\right)^2\right) + O(n^{-\frac{1}{2}})\,.$$

При помощи ортогональной матрицы вида

$$A = \begin{bmatrix} \sqrt{p_1} & \ldots & \sqrt{p_r} \\ a_{21} & \ldots & a_{2r} \\ \ldots & \ldots & \ldots \\ a_{r1} & \ldots & a_{rr} \end{bmatrix}$$

рассмотрим вектор $u = (u_1, \ldots, u_r)^T = At$, где $t = (t_1, \ldots, t_r)^T$, с первой компонентой $u_1 = \sum_k \sqrt{p_k}t_k$.

$$\sum_k t_k^2 = t^T t = t^T A^T A t = u^T u = \sum_k u_k^2,$$

$$Q(t_1,\dots,t_r)=\sum_k t_k^2-\left(\sum_k\sqrt{p_k}t_k\right)^2=\sum_k u_k^2-u_1^2=\sum_{k=2}^r u_k^2\,.$$

Следовательно, $Q(t_1,\dots,t_r)$ неотрицательна и имеет ранг $r-1$, и матрица вторых моментов имеет $r-1$ характеристических чисел, равных 1 и одно характеристическое число, равное 0.

Таким образом, при достаточно больших n совместная характеристическая функция $x_1,\dots,x_r$ имеет вид характеристической функции многомерного нормального закона с нулевым вектором средних и с матрицей вторых моментов ранга $r-1$, $\sum_k x_k\sqrt{p_k}=0$. Ортогональное преобразование переводит вектор $(x_1,\dots,x_r)$ в вектор с некоррелированными компонентами, у которого диагональная матрица вторых моментов имеет один ноль на главной диагонали. То есть одна из новых некоррелированных компонент имеет нулевую дисперсию, или, проще говоря, равна нулю. Сумма квадратов при ортогональном преобразовании равна сумме квадратов новых некоррелированных компонент, следовательно, $\sum_{k=1}^r x_k^2\sim\ \chi^2(r-1)$.

3.2.2 Случай неизвестных параметров

В случае, когда параметры распределения оцениваются по выборке, для важного класса оценок Фишером было предложено изменение, согласно которому *число степеней свободы нужно уменьшить на количество оцениваемых параметров.* Рассмотрим статистику

$$\chi^2=\sum_{i=1}^r\frac{(\nu_i-np_i(\alpha_1,\dots,\alpha_s))^2}{np_i(\alpha_1,\dots,\alpha_s)},\tag{18}$$

в которой p_i являются функциями от выборочных наблюдений, и свойства выборочного распределения зависят от метода оценивания параметров. Нейманом и Пирсоном было изучено предельное распределение величины χ^2 при оценке параметров по

методу минимума хи-квадрат. Относительно $\alpha_1, \ldots, \alpha_s$ решается система уравнений

$$-\frac{1}{2}\frac{\partial \chi^2}{\partial \alpha_j} = \sum_{i=1}^{r} \left(\frac{\nu_i - np_i}{p_i} + \frac{(\nu_i - np_i)^2}{2np_i^2} \right) \frac{\partial p_i}{\partial \alpha_j} = 0\,,$$

где $j = 1, 2, \ldots, s$. Полученные таким образом оценки α_j подставляются в (18). Если при больших n вторым слагаемым можно пренебречь, считая $2np_i^2$ постоянным, то получим систему

$$\sum_{i=1}^{r} \frac{\nu_i - np_i}{p_i} \frac{\partial p_i}{\partial \alpha_j} = 0\,. \tag{19}$$

Метод оценки параметров из этой системы называется *видоизмененным методом минимума хи-квадрат.*

Теорема 6 *Пусть заданы функции $p_i(\alpha_1, \ldots, \alpha_s)$ от $s < r$ переменных $\alpha_1, \ldots, \alpha_s$, $i = 1, 2, \ldots, r$, удовлетворяющие условиям:*

1. $\sum\limits_{i=1}^{r} p_i(\alpha_1, \ldots, \alpha_s) = 1$;
2. $p_i(\alpha_1, \ldots, \alpha_s) > c^2$;
3. *все p_i имеют непрерывные производные $\frac{\partial p_i}{\partial \alpha_j}$, $\frac{\partial^2 p_i}{\partial \alpha_j \partial \alpha_k}$;*
4. *матрица первых производных имеет ранг s.*

Тогда уравнения (19) имеют одно решение $\alpha = (\alpha_1, \ldots, \alpha_s)$, сходящееся по вероятности к α_0 при $n \to \infty$. Значение статистики (18) при таких α асимптотически распределено по $\chi^2(r - s - 1)$.

3.3 Проверка гипотезы согласия с нормальным распределением

На рис. 1 представлена гистограмма относительных частот выборки, смоделированной по нормальному закону с параметрами $\mu = 50$, $\sigma = 20$. Выясним, насколько согласовано эмпирическое распределение с нормальным $\mathcal{N}(\mu, \sigma)$. В качестве оценок рассмотрим $\hat{\mu} = \bar{x} = 48.72$, $\hat{\sigma} = S = 20.47$.

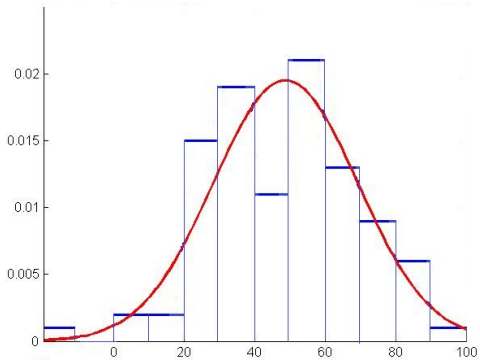

Рис. 1: Выборочная гистограмма и плотность $\mathcal{N}(\mu, \sigma)$.

Обозначим через $(z_{i-1}; z_i]$ интервал S_i, $i = 1, \ldots, r$, $z_0 = -\infty$, $z_r = +\infty$, через ν_i количество элементов выборки x_k, таких что $z_{i-1} < x_k \leq z_i$. Для вычисления вероятностей p_i воспользуемся функцией $\Phi(x)$ стадартного нормального распределения[2].

$$p_i = \Phi\left(\frac{z_i - \hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{z_{i-1} - \hat{\mu}}{\hat{\sigma}}\right). \tag{20}$$

Объединим наблюдения в крайних ячейках так, чтобы все np_i были не меньше пяти. Результаты вычислений представлены в следующей таблице.

[2]В электронных таблицах $\Phi(x)$ можно получить при помощи функции НОРМСТРАСП(x).

i	$(z_{i-1}; z_i]$	ν_i	$\Phi(\frac{z_i-\hat{\mu}}{\hat{\sigma}})$	p_i	$\frac{(\nu_i-np_i)^2}{np_i}$
1	$(-\infty; 20]$	5	0.080	0.080	1.144
2	$(20; 30]$	15	0.180	0.100	2.510
3	$(30; 40]$	19	0.335	0.155	0.799
4	$(40; 50]$	11	0.525	0.190	3.360
5	$(50; 60]$	21	0.709	0.184	0.359
6	$(60; 70]$	13	0.851	0.142	0.094
7	$(70; 80]$	9	0.937	0.086	0.018
8	$(80; +\infty)$	7	1	0.063	0.072

Суммируя элементы в последнем столбце, получаем значение статистики $\chi^2_* = 8.356$. Число степеней свободы равно $df = 8 - 1 - 2 = 5$.

Для проверки гипотезы о том, что случайная величина имеет определенный закон распределения, в множестве значений случайной величины мы выделяем критическую область, вероятность α попадания в которую настолько мала, что мы считаем это невозможным. Если случайная величина туда все-таки попала, то мы будем считать, что эта скорее имеет какое-то другое, а не предполагаемое распределение. Если случайная величина не попала в эту маловероятную (критическую) область, то условимся считать, что оснований отвергнуть гипотезу нет.

Критическое значение, соответствующее уровню значимости $\alpha = 0.05$, вычисляется при помощи специальных таблиц, вероятностных калькуляторов или электронных таблиц *Excel* и равно $\chi^2_{0.95} = \text{ХИ2ОБР}(0.05; 5) = 11.07$. Наблюдаемое значение статистики $\chi^2_* = 8.356$ меньше критического, следовательно, гипотеза о согласии эмпирического распределения с нормальным не отвергается с уровнем значимости $\alpha = 0.05$. Доверительный уровень вероятности можно вычислить

$$p = P\{\chi^2 > \chi^2_*\} = \text{ХИ2РАСП}(8.356; 5) = 0.14 > \alpha = 0.05.$$

При $p < \alpha$ мы бы сказали, что эмпирическое распределение плохо согласуется с нормальным.

3.4 Критерий хи-квадрат независимости (л.6)

Предположим, что у нас имеется n наблюдений, характеризующихся значениями двух качественных признаков X и Y с градациями $x_1, \ldots, x_r$ и $y_1, \ldots, y_s$.

ПРИМЕР 1.1. X – это случайный ответ на вопрос „У меня бывают периоды плача и близости к слезам“, допускающий четыре градации x_1 – „совершенно не согласен“, x_2 – „скорее не согласен“, x_3 – „скорее согласен“, x_4 – „совершенно согласен“. Признак Y указывает на пол: y_1 – женский, y_2 – мужской. Из $n = 36$ студентов, участвующих в анкетировании, 14 студентов „совершенно не согласны“, из них 6 девушек и 8 юношей, „скорее не согласны“ пятеро (4 девушки и 1 юноша), 7 человек „скорее согласны“ (1 юноша и 6 девушек) и 10 „совершенно согласны“ (один юноша и 9 девушек).

$X \setminus Y$	y_1	y_2	$\ldots$	y_s	сумма
x_1	n_{11}	n_{12}	$\ldots$	n_{1s}	n_{1*}
x_2	n_{21}	n_{22}	$\ldots$	n_{2s}	n_{2*}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	n_{r*}
сумма	n_{*1}	n_{*2}	$\ldots$	n_{*s}	n

$X \setminus Y$	y_1	y_2	сумма
x_1	6	8	14
x_2	4	1	5
x_3	6	1	7
x_4	9	1	10
сумма	25	11	36

Обозначим через

$$p_{ij} = P(X = x_i, Y = y_j),\ p_{i*} = P(X = x_i),\ p_{*j} = P(Y = y_j).$$

Если X и Y *независимы*, тогда для любых $i = 1, 2, \ldots, r$ и $j = 1, 2, \ldots, s$ справедливо

$$p_{ij} = p_{i*}p_{*j}. \tag{21}$$

При конечном количестве наблюдений оценками вероятностей являются отношения

$$\widehat{p}_{ij} = \frac{n_{ij}}{n}, \quad \widehat{p}_{i*} = \frac{n_{i*}}{n}, \quad \widehat{p}_{*j} = \frac{n_{*j}}{n}, \tag{22}$$

и крайне редко удается получить равенства вида $\frac{n_{ij}}{n} = \frac{n_{i*}}{n} \cdot \frac{n_{*j}}{n}$, например $\widehat{p}_{11} = \frac{6}{36} \neq \frac{14}{36} \cdot \frac{25}{36} = \widehat{p}_{1*}\widehat{p}_{*1}$. Важно определить, не является ли это нарушение случайным.

Для этого используется статистика χ^2 Пирсона (18), в выражении которой в качестве вероятностей гипотетического распределения рассматриваются произведения $\widehat{p}_{i*} \cdot \widehat{p}_{*j}$, соответствующие независимости признаков. Вероятности $\widehat{p}_{i*}$ и $\widehat{p}_{*j}$ оцениваются по выборке.

$$\widehat{p}_{i*} = \frac{n_{i*}}{n}, \quad \widehat{p}_{*j*} = \frac{n_{*j}}{n},$$

статистика χ^2 имеет вид

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - np_{ij})^2}{np_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - n\frac{n_{i*}}{n}\frac{n_{*j}}{n})^2}{n\frac{n_{i*}}{n}\frac{n_{*j}}{n}} =$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2 - 2n_{ij}\frac{n_{i*}n_{*j}}{n} + \left(\frac{n_{i*}n_{*j}}{n}\right)^2}{\frac{n_{i*}n_{*j}}{n}} =$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{s} \left(\frac{n \cdot n_{ij}^2}{n_{i*}n_{*j}} - 2n_{ij} + \frac{n_{i*}n_{*j}}{n}\right) = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i*}n_{*j}} - 1\right)$$

и при $n \to \infty$ имеет распределение хи-квадрат с числом степеней свободы, равным

$$rs - 1 - (r-1) - (s-1) = (r-1)(s-1)\,,$$

так как $\sum\limits_{i=1}^{r} p_{i*} = \sum\limits_{j=1}^{s} p_{*j} = 1$, и количество оцениваемых параметров по выборке равно $r + s - 2$.

Получаем значение статистики $\chi^2 = 7.79$ с числом степеней свободы, равным $(4-1)(2-1) = 3$. В электронных таблицах *Excel* для получения P-значения можно воспользоваться встроенной функцией $P = \text{ХИ2РАСП}(7.79; 3) = 0.05$. Следовательно, независимость признаков отвергается, и то, что есть юноши и девушки по-разному отвечают на поставленный вопрос.

При использовании этого критерия следует учитывать асимптотический характер распределения статистики и при небольших объемах выборки предпочтительнее обращаться к другим критериям, которые будем рассматривать далее.

В частном случае $r = 2$, $s = 2$ и частотах $n_{11} = a$, $n_{12} = b$, $n_{21} = c$, $n_{22} = d$ в качестве статистики хи-квадрат независимости используется выражение

$$\chi^2 = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}\,. \tag{23}$$

Действительно, $\frac{\chi^2}{n}$ имеет вид

$$\frac{a^2}{(a + b)(a + c)} + \frac{b^2}{(a + b)(b + d)} + \frac{c^2}{(a + c)(c + d)} + \frac{d^2}{(b + d)(c + d)} - 1\,.$$

Отдельно рассмотрим числитель этого выражения.

$$\begin{aligned}
&a^2(b + d)(c + d) + b^2(a + c)(c + d) + c^2(a + b)(b + d) + \\
&+d^2(a + b)(a + c) - (bc + cd + bd + d^2)(a^2 + ab + ac + bc) = \\
&\qquad\qquad \left\{ \begin{array}{c} a^2(bc + cd + bd + d^2) \\ +b^2(ac + c^2 + ad + cd) \\ +c^2(ab + b^2 + ad + bd) \\ +\mathbf{d^2}(\mathbf{a^2} + \mathbf{ab} + \mathbf{ac} + \mathbf{bc}) \end{array} \right\} - \\
&\qquad -(bc + cd + bd + \mathbf{d^2})(a^2 + ab + ac + bc) = \\
&\qquad\qquad \left\{ \begin{array}{c} \mathbf{a^2}(\mathbf{bc} + \mathbf{cd} + \mathbf{bd} + d^2) \\ +b^2(ac + c^2 + ad + cd) \\ +c^2(ab + b^2 + ad + bd) \end{array} \right\} - \\
&\qquad -(bc + cd + bd)(\mathbf{a^2} + ab + ac + bc) = \\
&\left\{ \begin{array}{c} a^2d^2 + \\ +b^2(ac + \mathbf{c^2} + ad + cd) \\ +c^2(ab + b^2 + ad + bd) \end{array} \right\} \quad - \quad \left\{ \begin{array}{c} bc(ab + ac + bc) \\ +cd(\mathbf{ab} + ac + bc) \\ +bd(ab + \mathbf{ac} + bc) \end{array} \right\} = \\
&\qquad = a^2d^2 - 2abcd + b^2c^2 = (ad - bc)^2\,.
\end{aligned}$$

4 Статистический анализ качественных признаков

4.1 Коэффициент неопределенности

Всем хорошо известно, что в любом живом существе происходит передача информации. Органы чувств собирают информацию о внешнем мире, нервная система передает ее в мозг, который ее перерабатывает и рассылает приказы на основе поступившей информации по нервным волокнам в органы и мышцы. Аналогично передается информация в любом действующем учреждении: передача информации осуществляется в виде докладных записок, распоряжений, запросов и т.д. Вычислительная техника по заданной программе, т.е. на основе заданной информации, с огромной скоростью перерабатывает большие количества информации. Перечень примеров можно было бы продолжить, но ясно одно, что уровень достижения в большей степени определяется оперативностью работы с информацией, т.е. тем, с какой скоростью и с какой надежностью осуществляется ее передача, хранение (защита) и переработка. Вопросы передачи и хранения информации тесно связаны со способом ее измерения и кодирования. Вопросы переработки информации опираются на линейные статистические методы.

Измерять информацию, подобно расстоянию, массе, количеству теплоты и т.д., стали относительно недавно. Поясним способ измерения информации на основе игры "Бар-Кохба". Один из участников игры должен отгадать, что задумал другой, задавая ему вопросы, допускающие только два ответа: "да"или "нет". Обычно задумывают какие-либо предметы или строки стихотворений. Количество информации, необходимое для отгадывания, можно измерить количеством вопросов, которые требуется задать при наиболее рациональной тактике дознания.

Запишем полученные ответы, заменив единицами положительные ответы и нулями отрицательные. Получим последова-

тельность нулей и единиц

$$001100011010\ldots,$$

заменяющую последовательность ответов. Процесс замены ответов нулями и единицами называется *кодированием*, а сама последовательность – *кодовым словом*. Длина кодового слова может служить мерой количества информации.

Ответ на любой вопрос, допускающий два равно вероятных ответа "да" или "нет", содержит одну единицу информации, которую называют бит.

Иными словами, количество информации, которое можно с одинаковой вероятностью закодировать нулем или единицей, будем считать единицей информации. Заметим, что при измерении количества информации мы сознательно оставляем без внимания содержание информации.

4.1.1 Формула Хартли

Количество информации H, необходимое для выявления объекта, о котором только известно, что он принадлежит множеству из N объектов, вычисляется при помощи *формулы Хартли*:

$$H = log_2 N. \tag{24}$$

Например, известно, что у одного из 16 студентов сегодня день рождения. Эту неопределенность можно устранить, задав несколько вопросов тому, кто знает. Сколько вопросов нужно задать, если ответы могут быть только "да" или "нет"? – Ответ: $4 = log_2 16$. Для этого лучше всего разделить группу на две части по восемь человек и задать вопрос: "В первой половине счастливчик?". Восьмерку со счастливчиком следует опять разделить пополам и задать аналогичный вопрос, и т.д. Четвертый вопрос устранит неопределенность полностью.

В случае, когда количество N элементов множества не является степенью двойки, количество информации, вычисляемое по формуле Хартли, не является целым, например, $log_2 12 = 3.58$.

4.1.2 Формула Шеннона

Если в множестве элементы не равновероятны, то для вычисления информации, необходимой для отгадывания элемента этого множества, используется *формула Шеннона.* Обозначим через

$$\xi : \begin{pmatrix} x_1 & \dots & x_n \\ p_1 & \dots & p_n \end{pmatrix}$$

случайный эксперимент, а через $x_1, \dots, x_n$ возможные исходы этого эксперимента, которые осуществляются с вероятностями $p_1, \dots, p_n$, $\sum\limits_{k=1}^{n} p_k = 1$. Количество информации $H(\xi)$, которую нужно получить, чтобы узнать, какой исход эксперимента осуществлен, вычисляется по формуле Шеннона:

$$H(\xi) = \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k}. \tag{25}$$

Например, из перемешанной колоды 36 карт вынимается одна карта. Значениями случайного эксперимента ξ_1 являются все 36 карт, вероятность появления каждой равна $1/36$:

$$\xi_1 : \begin{pmatrix} x_1 & \dots & x_{36} \\ \frac{1}{36} & \dots & \frac{1}{36} \end{pmatrix}.$$

В этом случае формула Шеннона совпадает с формулой Хартли

$$H(\xi_1) = \sum_{k=1}^{36} \frac{1}{36} \log_2 36 = \log_2 36 = 5.17.$$

Пусть ξ_2 – некий случайный эксперимент, в результате которого из такой же колоды карт вынимается или нет туз:

$$\xi_2 : \begin{pmatrix} \text{'не туз'} & \text{'туз'} \\ \frac{8}{9} & \frac{1}{9} \end{pmatrix}$$

$$H(\xi_2) = \frac{8}{9} \log_2 \frac{9}{8} + \frac{1}{9} \log_2 \frac{9}{1} = 0.352 + 0.151 = 0.503.$$

Необходимая для отгадывания результата случайного эксперимента ξ информация $H(\xi)$ равна неопределенности этого эксперимента, поэтому часто $H(\xi)$ называют *энтропией* распределения случайного эксперимента ξ.

Пусть ξ_0 – некий случайный эксперимент, в результате которого из такой же колоды карт вынимается красная или черная карта. Имеем два равновероятных исхода, поэтому должны получить один бит информации:

$$\xi_0 : \begin{pmatrix} \text{'красная'} & \text{'черная'} \\ 0.5 & 0.5 \end{pmatrix}$$

$$H(\xi_0) = 0.5 \log_2 2 + 0.5 \log_2 2 = 0.5 + 0.5 = 1.$$

4.1.3 Условная энтропия и количество информации

Пусть имеются случайные эксперименты

$$\xi : \begin{pmatrix} A_1 & \dots & A_N \\ p_1 & \dots & p_N \end{pmatrix} \quad \text{и} \quad \eta : \begin{pmatrix} B_1 & \dots & B_n \\ q_1 & \dots & q_n \end{pmatrix}.$$

Каждое событие B_i изменяет вероятности исходов $A_1, \dots, A_N$. Обозначим через $p_j(B_i) = P(A_j|B_i)$ соответствующие условные вероятности и определим *условную энтропию*

$$H_{B_i}(\xi) = -\sum_{j=1}^{N} p_j(B_i) \log_2 p_j(B_i)\,. \tag{26}$$

Средняя условная энтропия $H_\eta(\xi)$ случайного эксперимента ξ при известном исходе эксперимента η равна

$$H_\eta(\xi) = H(\xi|\eta) = q_1 H_{B_1}(\xi) + \dots q_n H_{B_n}(\xi). \tag{27}$$

Разность между исходной энтропией $H(\xi)$ и средней условной энтропией $H_\eta(\xi) = H(\xi|\eta)$ называется *количеством информации* $I(\xi, \eta)$ о случайном эксперименте ξ, полученном посредством η,

$$I(\xi, \eta) = H(\xi) - H(\xi|\eta). \tag{28}$$

Убедимся в том, что $H(\eta|\xi) = H(\xi,\eta) - H(\xi)$. Действительно,

$$H(\eta|\xi) = \sum_{k=1}^{r} p_{k\cdot} \left(-\sum_{j=1}^{s} \frac{p_{kj}}{p_{k\cdot}} (\log_2 p_{kj} - \log_2 p_{k\cdot}) \right) =$$
$$= -\sum_{k=1}^{r}\sum_{j=1}^{s} p_{kj} \log_2 p_{kj} + \sum_{k=1}^{r}\sum_{j=1}^{s} p_{kj} \log_2 p_{k\cdot} =$$
$$= H(\xi,\eta) + \sum_{k=1}^{r} \log_2 p_{k\cdot} \sum_{j=1}^{s} p_{kj} = H(\xi,\eta) - H(\xi).$$

Отсюда

$$I(\xi,\eta) = H(\xi) + H(\eta) - H(\xi,\eta). \tag{29}$$

Для независимых случайных экспериментов ξ и η количество информации равно нулю, так как $\frac{p_{kj}}{p_{k\cdot}} = p_{\cdot j}$ для любого x_k, $H(\eta|\xi = x_k) = H(\eta)$ и $I(\xi,\eta) = 0$.

С помощью *односторонних коэффициентов неопределенности*

$$J_{X|Y} = \frac{I(X,Y)}{H(Y)} \cdot 100\% \quad \text{и} \quad J_{Y|X} = \frac{I(X,Y)}{H(X)} \cdot 100\% \tag{30}$$

измеряется доля информации одного признака, которую он разделяет с другим. Если интерес представляет не только то, как Y влияет на X, но и то, как сам X влияет на Y, то используют *двусторонний или симметричный коэффициент неопределенности*, равный

$$J = \frac{H(X)}{H(X)+H(Y)} J_{X|Y} + \frac{H(Y)}{H(X)+H(Y)} J_{Y|X} =$$
$$= \frac{2I(X,Y)}{H(X)+H(Y)} \cdot 100\%\,. \tag{31}$$

ПРИМЕР. Из биологии известно, что за языковые способности отвечает левое полушарие. Рассмотрим таблицу сопряженности двух призна-

ков: X — наличие у больного глиобластомы в правом или в левом полушарии головного мозга; Y — наблюдаемые нарушения речи.

	нарушения речи		
Полушарие	нет	есть	всего
правое	155	12	167
левое	94	59	153
всего	249	71	320

$H(X) = 0.9986$, $H(Y) = 0.7636$, $H(X,Y) = 1.6531$, $I(X,Y) = 0.11$. Коэффициент $J_{X|Y} = 11\%$ указывает на то, какую часть информации о том, какое полушарие повреждено, мы получаем, если узнаем о том, имеются у больного нарушения речи или нет. При отсутствии речевых нарушений у 62 процентов больных поражено правое полушарие, а при нарушении речи у 83 процентов больных глиобластома в левом полушарии. Коэффициент $J_{Y|X} = 14\%$ указывает на то, какую часть информации о том, имеются или нет нарушения речи, мы получаем, если знаем, какое полушарие повреждено: речевые нарушения встречаются у 8 процентов больных с глиобластомой в правом и у 39 процентов больных с глиобластомой в левом полушарии.

4.2 Точный критерий Фишера

Точный критерий Фишера используется для проверки гипотезы однородности категориальных вероятностей в случае таблиц сопряженности размерности 2×2.

ПРИМЕР 1.2. Рассмотрим результаты некоторого наблюдения, согласно которому в 5 из 21 страны уровень сметности от заболеваний системы кровообращения (СК) очень высокий, причем в четырех из них высокий уровень смертности от заболеваний сосудов головного мозга (СГМ). Во странах с высокой смертностью от заболеваний СГМ высокая смертность от заболеваний СК. Составляем таблицу сопряженности по двум признакам: X — фактор смертности от заболеваний СК: $X = 0$ при низкой и $X = 1$ при высокой; Y — фактор смертности от заболеваний СГМ: $Y = 0$ также при низкой и $Y = 1$ при высокой.

$X \setminus Y$	0	1	
0	16	0	16
1	1	4	5
	17	4	21

$X \setminus Y$	0	1	
0	a	b	$a+b$
1	c	d	$c+d$
	$a+c$	$b+d$	n

Обозначим через

$$p_0 = P\{Y = 1|X = 0\} \text{ и } p_1 = P\{Y = 1|X = 1\} \tag{32}$$

вероятности высокой смертности от ЗСГМ. Проверяемая гипотеза имеет вид $H_0 : p_0 = p_1$ означает, что вероятность высокой смерности от ЗСГМ не зависит от того, какая смерность от ЗСК. Вычислим вероятность P^a_{a+c} того, что из $a + c$ стран с низкой смертностью от ЗСГМ в a странах оказалась высокая смертность от ЗСК.

Напомним, что N-множеством в комбинаторике называется множество, состоящее из N элементов. Его K-подмножество называется сочетанием. Упорядоченное K-подмножество — размещением. Число размещений из N элементов по K обозначают через A^K_N. Оно равно произведению из K убывающих сомножителей

$$A^K_N = N \cdot (N - 1) \cdot \ldots \cdot (N - k + 1),$$

так как на первое место упорядоченного K-множества можно выбрать любой элемент N-множества, на второе — любой из оставшихся $(N - 1)$ элементов, и так далее до выбора на K-е место любого из $(N - K + 1)$ элементов. Очевидно, что

$$A^K_N = \frac{N!}{(N - K)!}\,.$$

В частности, $A^N_N = N!$. Число сочетаний обозначается через C^K_N. Оно связано с числом размещений выражением $C^K_N \cdot K! = A^K_N$, поэтому для вычисления числа сочетаний используют формулу

$$C^K_N = \frac{N!}{K!(N - K)!}. \tag{33}$$

Для вычисления вероятности P^a_{a+c} сначала требуется вычислить число исходов, удовлетворяющих тому, что $a + c$ стран из

$n = a+b+c+d$ имеют низкую сметрность от ЗСГМ. Это число равно

$$C_n^{a+c} = \frac{n!}{(a+c)!(b+d)!}.$$

Затем вычислим число исходов, удовлетворяющих тому, что в a странах из $a+c$ сметрность от ЗСК высокая, а в c странах низкая. При справедливости нулевой гипотезы $H_0 : p_0 = p_1$ это число исходов равно произведению

$$C_{a+b}^{a} \cdot C_{c+d}^{c} = \frac{(a+b)!(c+d)!}{a!b!c!d!},$$
$$P_{a+c}^{a} = \frac{C_{a+b}^{a} \cdot C_{c+d}^{c}}{C_n^{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}.$$

При заданных маргинальных частотах $a+b$, $c+d$, $a+c$, $b+d$ можно перечислить все возможные таблицы с положительными элементами вида

$T \sim$

$X \setminus Y$	0	1	
0	x	$a+b-x$	$a+b$
1	$a+c-x$	$(c+d)-(a+c-x)$	$c+d$
	$a+c$	$b+d$	$a+b+c+d$

и вычислить вероятность получения этой таблицы по формуле

$$P(T) = \frac{(a+b)!(a+c)!(c+d)!(b+d)!}{x!(a+b-x)!(a+c-x)!(d-a+x)!(a+b+c+d)!}.$$

Для каждой из возможных таблиц сопряженности вычисляется характеристика, отражающая степень отклонения от так называемой идеальной таблицы, соответствующей независимым признакам. Для примера будем рассматривать статистику хи-квадрат

$$\chi^2(T_0) = \frac{(ad-bc)^2 \cdot (a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)}.$$

Значимость критерия Фишера по таблице T_0 вычисляется как вероятность получить двумерное распределение, при котором зависимость между признаками выражена не меньше, чем в случае распределения, соответствующего T_0.

$$\alpha_* = \sum P(T|\chi^2(T) \geq \chi^2(T_0))\,.$$

На основе принципа маловероятных событий при $\alpha_* < \alpha$, где α — уровень значимости, нулевая гипотеза отвергается.

4.3 Критерий для проверки равенства частот

Пусть имеется парная выборка $(x_1, y_1), \ldots, (x_n, y_n)$ наблюдений над бернуллиевскими случайными величинами X и Y. Обозначим через $p = P\{X = 1\}$ вероятность успеха для X. Предположим независимость X и Y. Тогда будут одинаковыми условные вероятности успеха $p_1 = p_2 = p$ из (32). Очевидно,

$$\mathbf{E}\bar{X} = p, \mathbf{D}\bar{X} = p(1-p).$$

Представим наблюдения в виде таблицы сопряженности

$X \setminus Y$	0	1	сумма
0	a	b	$a+b$
1	c	d	$c+d$
сумма	$a+c$	$b+d$	n

Предложение 6 *Пусть справедлива нулевая гипотеза*

$$H_0 : p_1 = p_2 = p. \tag{34}$$

Тогда стандартное нормальное распределение имеет статистика вида

$$Z = \frac{(bc - ad)\sqrt{n}}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \sim \mathcal{N}(0,1). \tag{35}$$

Доказательство. В качестве оценок будем рассматривать

$$\hat{p}_1 = \frac{c}{a+c}, \quad \hat{p}_2 = \frac{d}{b+d}, \quad \hat{p} = \frac{c+d}{n}, \quad \text{где} \quad n = a+b+c+d.$$

Заметим, что $\hat{p}_1$ и $\hat{p}_1$ являются выборочными средними выборки X при условиях $Y = 0$ и $Y = 1$, т.е. $\hat{p}_1 = \bar{x}_1$ и $\hat{p}_2 = \bar{x}_2$. С учетом предположения (41) их математические ожидания равны

$$\mathbf{E}\bar{x}_1 = \mathbf{E}\bar{x}_2 = p, \ \Rightarrow \ \mathbf{E}(\bar{x}_1 - \bar{x}_2) = 0$$

а дисперсии на основе независимости элементов выборки имеют вид:

$$\mathbf{D}\bar{x}_1 = \frac{p(1-p)}{a+c}, \qquad \mathbf{D}\bar{x}_2 = \frac{p(1-p)}{b+d}, \ \Rightarrow$$
$$D = \mathbf{D}(\bar{x}_1 - \bar{x}_2) = \frac{p(1-p)}{\frac{1}{a+c} + \frac{1}{b+d}}.$$

Согласно ЦПТ,

$$\bar{x}_1 - \bar{x}_2 = \frac{c}{a+c} - \frac{d}{b+d} = \frac{bc-ad}{(a+c)(b+d)} \sim \mathcal{N}(0, \sqrt{D}),$$

следовательно, при справедливости нулевой гипотезы (41) статистика $Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{D}}$ имеет стандартное нормальное распределение. При $\hat{p} = \frac{c+d}{n}$ алгебраически приходим к (35) следующим образом:

$$Z = \frac{\frac{bc-ad}{(a+c)(b+d)}}{\sqrt{\frac{(c+d)}{n}\frac{(a+b)}{n}\left(\frac{1}{a+c} + \frac{1}{b+d}\right)}} = \frac{(bc-ad)\sqrt{n}\sqrt{(a+c)(b+d)}}{(a+c)(b+d)\sqrt{(c+d)(a+b)}}.$$

Замечание. Квадрат статистики Z в точности совпадает со значением статистики хи-квадрат (23) независимости признаков с двумя градациями.

Если вычислить коэффициент корреляции между признаками X и Y по формуле

$$R = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_i x_i^2 - n\bar{x}^2}\sqrt{\sum_i y_i^2 - n\bar{y}^2}},$$

то используя то, что

$$\sum_i x_i y_i = d, \quad \sum_i x_i^2 = c + d, \quad \sum_i y_i^2 = b + d,$$
$$\bar{x} = \frac{c+d}{n}, \quad \bar{y} = \frac{b+d}{n},$$

можно получить выражение

$$R = \frac{d - (c+d)(b+d)/n}{\sqrt{(c+d) - (c+d)^2/n}\sqrt{(b+d) - (b+d)^2/n}} =$$
$$= \frac{d(a+b+c+d) - (c+d)(b+d)}{\sqrt{(c+d)(b+d)}\sqrt{n - (c+d)}\sqrt{n - (b+d)}} =$$
$$= \frac{ad - bc}{\sqrt{(c+d)(b+d)(a+b)(a+c)}}.$$

1. Влияние смертности от несчастных случаев и других внешних воздействий (Y) на естественный прирост населения в 1997 году.

ест.прирост в 1997	$Y < 60$	$Y \geq 60$	сумма
низк.	3	4	7
выс.	10	4	14
сумма	13	8	21

2. Влияние смертности от злокачественных новообразований (Y) на естественный прирост населения в 1997 году.

ест.прирост в 1997	$Y < 180$	$Y \geq 180$	сумма
низк.	3	4	7
выс.	3	11	14
сумма	6	15	21

3. Влияние смертности от ишемической болезни сердца (Y) на естественный прирост населения в 1997 году.

ест.прирост в 1997	$Y < 150$	$Y \geq 150$	сумма
низк.	1	6	7
выс.	7	7	14
сумма	8	13	21

4. Влияние смертности от заболеваний сосудов головного мозга (Y) на естественный прирост населения в 1997 году.

ест.прирост в 1997	$Y < 190$	$Y \geq 100$	сумма
низк.	3	4	7
выс.	14	0	14
сумма	17	4	21

4.4 Таблицы сопряженности для зависимых выборок (л.7)

Рассмотрим данные такого рода: X_1 и X_2— наличие в ЭКГ эпизода ускоренного идиовентрикулярного ритма (УИР) до и после операции АКШ. У 57 больных этот эпизод не наблюдался ни до, ни после операции; у восьми больных он был только до операции, у двух больных появился после. Не было ни одного больного, у которого эпизод проявлял устойчивость. Проверяется гипотеза о том, что ситуации улучшения или ухудшения равновероятны. Имеется таблица сопряженности в виде матрицы 2×2

$X_1 \setminus X_2$	$-$	$+$	сумма
$-$	a	b	$a+b$
$+$	c	d	$c+d$
сумма	$a+c$	$b+d$	n

$X_1 \setminus X_2$	$-$	$+$	сумма
$-$	57	2	59
$+$	8	0	8
сумма	65	2	67

Нас интересует, насколько значимо различие между частотами b и c. Точная статистика критерия Мак Немара (McNemar's test)

вычисляется как

$$\alpha_* = 2 \sum_{i=1}^{\min(b,c)} C_{b+c}^i \frac{1}{2^{b+c}}\,.$$

При малых значениях $\alpha_* < \alpha = 0.05$ гипотеза о равенстве $b = c$ отвергается и различие между ними нельзя объяснить случайностью. В нашем случае $b = 2$, $c = 8$, $\alpha_* = 2(C_{10}^0 + C_{10}^1 + C_{10}^2)/2^{10} = 0.0429 < 0.05$, поэтому число больных, у которых исчез эпизод УИР после операции, значимо больше тех, у кого этот эпизод в ЭКГ появился.

Помимо этой статистики используется выражение

$$\chi^2 = \frac{(b-(b+c)/2)^2}{(b+c)/2} + \frac{(c-(b+c)/2)^2}{(b+c)/2} = \frac{(b-c)^2}{b+c},$$

в котором добавляется поправка на непрерывность

$$\chi_*^2 = \frac{(|b-c|-1)^2}{b+c}.$$

При справедливости нулевой гипотезы статистика χ^2 имеет распределение хи-квадрат с одной степенью свободы. Доверительный уровень вероятности равен $p = P\{\chi^2 > \chi_*^2 = 3.6\} = 0.058$, следовательно, различие между c и b нельзя объяснить случайностью с уровнем значимости, большим 0.058. С поправкой на непрерывность $p = P\{\chi^2 > \chi_*^2 = 2.5\} = 0.11$.

Для категориальных данных, повторяющихся многократно, используется обобщение критерия Мак Немара в виде критерия Кохрена (Cochren's Q test). Пусть имеется s дихотомических признаков у n индивидов. Для определенности закодируем нулем ответы „нет“ и единицей ответы „да“, количество положительных ответов у i-го индивида обозначим через x_{i*}, в j-й

момент через x_{*j}, $N = \sum\limits_{j=1}^{s} x_{*j} = \sum\limits_{i=1}^{n} x_{i*}$.

	X_1	X_2	...	X_s	сумма
1	x_{11}	x_{12}	...	x_{1s}	x_{1*}
2	x_{21}	x_{22}	...	x_{2s}	x_{2*}
...	...	...	...	...	...
n	x_{n1}	x_{n2}	...	x_{ns}	x_{n*}
сумма	x_{*1}	x_{*2}	...	x_{*s}	N

При отсутствии изменений в динамике наблюдений статистика

$$T = s(s-1)\frac{\sum\limits_{j=1}^{s}(x_{*j} - \frac{N}{s})^2}{\sum\limits_{i=1}^{n} x_{i*}(s - x_{i*})} \tag{36}$$

имеет распределение хи-квадрат с $s-1$ степенями свободы. Например, исследуется динамика появления эпизода УИР у $n = 52$ больных в $s = 4$ точках: до операции, через две недели после операции, через полгода и через год. Эпизод УИР появлялся у $N = 13$ больных однажды: $x_{j*} \in \{0, 1\}$, до операции у $x_{*1} = 6$ больных, после у $x_{*2} = 2$, затем $x_{*3} = 3$, $x_{*4} = 2$. Значение статистики (36) равно $\chi_* = 3.3$ со значимостью $p = P\{\chi^2 > \chi^2_*\} = 0.35$. Это свидетельствует о том, что снижение числа эпизодов УИР после операции можно объяснить случайностью.

5 Точные выборочные распределения

5.1 Распределение Стьюдента

Пусть $\xi, \xi_1, \ldots, \xi_n$ независимые нормально распределенные величины $\mathcal{N}(0, \sigma)$. Случайная величина вида

$$\zeta = \frac{\xi}{\sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}\xi_i^2}} = \frac{\xi}{\eta}, \tag{37}$$

имеет распределение *Стьюдента с n степенями свободы.*

Не умаляя общности, будем считать $\sigma = 1$. Тогда $\sum\limits_{i=1}^{n}\xi_i^2$ имеет распределение $\chi^2(n)$ с плотностью

$$\kappa_n(x) = \frac{2^{-\frac{n}{2}}}{\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}\,. \tag{38}$$

$$F_{a\xi}(x) = P\{a\xi < x\} = P\left\{\xi < \frac{x}{a}\right\} = F_\xi\left(\frac{x}{a}\right),$$
$$F_{\sqrt{\xi}}(x) = P\left\{\sqrt{\xi} < x\right\} = P\{\xi < x^2\} = F_\xi(x^2), \implies$$
$$f_{a\xi}(x) = \frac{1}{a} f_\xi\left(\frac{x}{a}\right), \qquad f_{\sqrt{\xi}}(x) = 2x f_\xi(x^2).$$

Для величины $\frac{1}{n}\sum\limits_{i=1}^{n}\xi_i^2$ получаем плотность $n\kappa_n(nx)$, а для величины $\eta = \sqrt{\frac{1}{n}\sum\limits_{i=1}^{n}\xi_i^2}$ плотность $2nx\kappa_n(nx^2)$. Совместное распределение ξ и η имеет вид:

$$\frac{2n\eta}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\left(n\eta^2\right)^{\frac{n}{2}-1} e^{-\frac{n\eta^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}} = c_n \eta^{n-1} e^{-\frac{\xi^2+n\eta^2}{2}},$$

$$\text{где} \quad c_n = \frac{\sqrt{2}\left(\frac{n}{2}\right)^{\frac{n}{2}}}{\sqrt{\pi}\Gamma(\frac{n}{2})}\,.$$

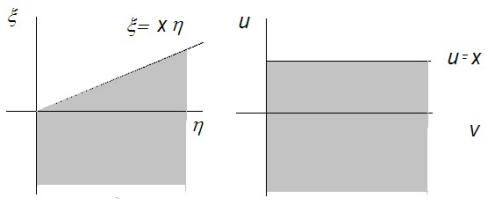

Рис. 2: Замена переменных $\xi = uv$, $\eta = v$ ($\eta > 0$,$\xi < x\eta$).

$$P\{\zeta < x\} = P\left\{\frac{\xi}{\eta} < x\right\} = P\{\xi < x\eta\} =$$

$$= c_n \iint\limits_{\substack{\eta>0\\ \xi<x\eta}} \eta^{n-1} e^{-\frac{\xi^2+n\eta^2}{2}} d\xi d\eta = c_n \int\limits_{-\infty}^{x} du \int\limits_{0}^{\infty} v^{n-1} e^{-\frac{u^2v^2+nv^2}{2}} v dv,$$

используем замену $\xi = uv$, $\eta = v$ с якобианом преобразования

$$J = \begin{vmatrix} \frac{\partial\xi}{\partial u} & \frac{\partial\xi}{\partial v} \\ \frac{\partial\eta}{\partial u} & \frac{\partial\eta}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v\,.$$

Далее заменим $\frac{(u^2+n)v^2}{2} = t$, $v = \frac{\sqrt{2t}}{\sqrt{u^2+n}}$, $vdv = \frac{dt}{u^2+n}$.

$$P\{\zeta < x\} = c_n \int\limits_{-\infty}^{x} du \int\limits_{0}^{\infty} \left(\frac{\sqrt{2t}}{\sqrt{u^2+n}}\right)^{n-1} e^{-t} \frac{dt}{u^2+n} =$$

$$= c_n 2^{\frac{n-1}{2}} \int\limits_{-\infty}^{x} \frac{du}{(u^2+n)^{\frac{n+1}{2}}} \Gamma\left(\frac{n+1}{2}\right) =$$

$$= \frac{\sqrt{2}\left(\frac{n}{2}\right)^{\frac{n}{2}}}{\sqrt{\pi}\Gamma(\frac{n}{2})} 2^{\frac{n-1}{2}} \int\limits_{-\infty}^{x} \frac{du}{(u^2+n)^{\frac{n+1}{2}}} \Gamma\left(\frac{n+1}{2}\right)$$

Отсюда плотность распределения имеет вид $f_\zeta(x,n) =$

$$= \frac{n^{\frac{n}{2}+\frac{1}{2}}}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(\frac{n}{2})} \frac{1}{(x^2+n)^{\frac{n+1}{2}}} = \frac{1}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(\frac{n}{2})} \left(\frac{x^2}{n}+1\right)^{-\frac{n+1}{2}}.$$

5.2 Асимптотическое свойство распределения Стьюдента

Плотность распределения Стьюдента

$$f_\zeta(x,n) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(\frac{n}{2})} \left(\frac{x^2}{n}+1\right)^{-\frac{n+1}{2}} . \tag{39}$$

можно записать в виде

$$f_\zeta(x,n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\frac{n}{2}}\Gamma(\frac{n}{2})} \frac{1}{\sqrt{2\pi}} \left(\frac{x^2}{n}+1\right)^{-\frac{n+1}{2}} .$$

Воспользуемся формулой Стирлинга

$$\ln\Gamma(\lambda) = \left(\lambda - \frac{1}{2}\right)\ln\lambda - \lambda + \frac{1}{2}\ln(2\pi) + \frac{1}{12\lambda} + O\left(\frac{1}{\lambda^3}\right).$$

При $p = \frac{n}{2} \longrightarrow \infty$ имеют место выражения

$$\begin{aligned} \ln\Gamma\left(p+\frac{1}{2}\right) - \frac{1}{2}\ln p - \ln\Gamma(p) \approx \\ \approx \left(p+\frac{1}{2}-\frac{1}{2}\right)\ln\left(p+\frac{1}{2}\right) - p - \frac{1}{2} - \\ -\frac{1}{2}\ln p - \left(p-\frac{1}{2}\right)\ln p + p = \\ = p\ln\left(p+\frac{1}{2}\right) - \frac{1}{2} - p\ln p = p\ln\left(1+\frac{1}{2p}\right) - \frac{1}{2} \approx 0 . \end{aligned}$$

При любом фиксированном x имеем

$$-\frac{n+1}{2}\ln\left(1+\frac{x^2}{n}\right) \underset{n\longrightarrow\infty}{\longrightarrow} -\frac{x^2}{2} ,$$

откуда

$$f_\zeta(x,n) \underset{n\longrightarrow\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} .$$

5.3 Распределение Фишера (л.8)

Распределенная по Фишеру случайная величина имеет вид

$$\zeta = \frac{\frac{1}{n}\sum\limits_{i=1}^{n}\xi_i^2}{\frac{1}{m}\sum\limits_{j=1}^{m}\eta_j^2} = \frac{m}{n}\zeta_0\,, \tag{40}$$

где ξ_i, η_j независимы, одинаково распределены $\mathcal{N}(0,\sigma)$. Не умаляя общности, будем считать $\sigma^2 = 1$. Тогда $\xi = \sum\limits_{i=1}^{n}\xi_i^2$ и $\eta = \sum\limits_{j=1}^{m}\eta_j^2$ имеют распределение (38) хи-квадрат.

$$P\{\zeta_0 < x\} = P\left\{\frac{\xi}{\eta} < x\right\} = P\{\xi < x\eta\}\,.$$

Совместное распределение величин ξ и η имеет вид

$$\kappa_n(\xi)\kappa_m(\eta) = \frac{2^{-\frac{n}{2}}}{\Gamma(\frac{n}{2})}\xi^{\frac{n}{2}-1}e^{-\frac{\xi}{2}}\frac{2^{-\frac{m}{2}}}{\Gamma(\frac{m}{2})}\eta^{\frac{m}{2}-1}e^{-\frac{\eta}{2}} =$$

$$= a_{nm}\xi^{\frac{n}{2}-1}\eta^{\frac{m}{2}-1}e^{-\frac{\xi+\eta}{2}},$$

$$\text{где}\quad a_{nm} = \frac{2^{-\frac{m+n}{2}}}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}\,.$$

$$P\{\zeta_0 < x\} = a_{nm}\int\limits_{\substack{\eta>0\\0<\xi<x\eta}}\xi^{\frac{n}{2}-1}\eta^{\frac{m}{2}-1}e^{-\frac{\xi+\eta}{2}}d\xi d\eta\,. \tag{41}$$

Используем замену $\xi = uv$, $\eta = v$ с якобианом преобразования

$$J = \begin{vmatrix} \frac{\partial\xi}{\partial u} & \frac{\partial\xi}{\partial v} \\ \frac{\partial\eta}{\partial u} & \frac{\partial\eta}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v\,.$$

Область интегрирования из $\eta > 0$, $0 < \xi < x\eta$ перейдет в область $v > 0, 0 < u < x$ (рис.3). Следовательно,

$$P\{\zeta_0 < x\} = a_{nm}\int\limits_0^x u^{\frac{n}{2}-1}du\int\limits_0^\infty v^{\frac{n+m}{2}-2}e^{-\frac{uv+v}{2}}vdv =$$

$$= a_{nm}\int\limits_0^x u^{\frac{n}{2}-1}du\int\limits_0^\infty v^{\frac{n+m}{2}-1}e^{-\frac{(u+1)v}{2}}dv\,.$$

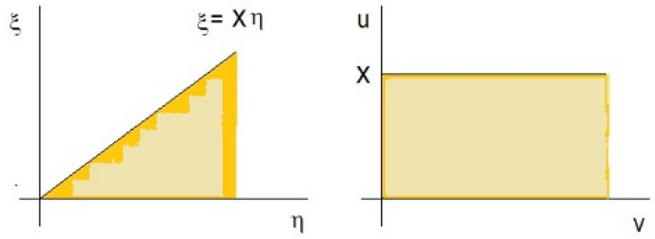

Рис. 3: Замена переменных $\xi = uv$, $\eta = v$.

Заменим переменные $t = \frac{u+1}{2}v,\ v = \frac{2}{u+1}$, $dv = \frac{2}{u+1}dt$.

$$a_{nm}\int\limits_0^x u^{\frac{n}{2}-1}du\int\limits_0^\infty t^{\frac{n+m}{2}-1}\left(\frac{2}{u+1}\right)^{\frac{n+m}{2}-1}\frac{2}{u+1}e^{-t}dt =$$

$$= \frac{2^{-\frac{m+n}{2}}}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}\int\limits_0^x u^{\frac{n}{2}-1}\left(\frac{2}{u+1}\right)^{\frac{n+m}{2}}du\Gamma\left(\frac{n+m}{2}\right) =$$

$$= \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}\int\limits_0^x \frac{u^{\frac{n}{2}-1}}{(u+1)^{\frac{n+m}{2}}}du\,.$$

Таким образом, используя $C_{nm} = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}$, получаем

$$f_0(x) = C_{nm}\frac{x^{\frac{n}{2}-1}}{(x+1)^{\frac{n+m}{2}}},$$

$$f(x) = \frac{n}{m}f_0\left(\frac{n}{m}x\right) = \frac{n}{m}C_{nm}\frac{\left(\frac{n}{m}x\right)^{\frac{n}{2}-1}}{\left(\frac{n}{m}x+1\right)^{\frac{n+m}{2}}}\,.$$

5.4 Распределение $\bar{x}$ и m_2 (л.8)

Лемма 1 *(Фишера) Пусть $x_1, \ldots, x_n$ — независимые и нормально распределенные $\mathcal{N}(0, \sigma)$ компоненты вектора X. В результате ортогонального преобразования $Y = CX$ имеем также независимые и нормально распределенные $\mathcal{N}(0, \sigma)$ компоненты $y_1, \ldots, y_n$. Тогда квадратичная форма*

$$Q(X) = \sum_{i=1}^{n} x_i^2 - y_1^2 - \ldots - y_p^2$$

независима от $y_1, \ldots, y_p$ и имеет плотность $\frac{1}{\sigma^2}\kappa_{n-p}\left(\frac{x}{\sigma^2}\right)$, где $\kappa_n(x)$ — плотность распределения хи-квадрат с n степенями свободы (38).

Теорема 7 *Пусть $x_1, \ldots, x_n \sim \mathbf{N}(\mu, \sigma)$, $\bar{x}$ и m_2 – выборочные среднее и второй центральный момент. Тогда*

1. *$\bar{x}$ и m_2 независимы;*
2. *$\bar{x} \sim \mathbf{N}(\mu; \frac{\sigma}{\sqrt{n}})$;*
3. *статистика $\frac{nm_2}{\sigma^2} \sim \boldsymbol{\chi^2(n-1)}$ имеет распределение хи-квадрат с $(n-1)$ степенью свободы.*

Без ограничения общности будем считать $\mu = 0$.

$$nm_2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

Выражение $n\bar{x}^2 = \left(\frac{x_1}{\sqrt{n}} + \ldots + \frac{x_n}{\sqrt{n}}\right)^2$ есть квадрат линейной формы $c_1x_1 + \ldots + c_nx_n$, для которой $c_1^2 + \ldots + c_n^2 = 1$. Поэтому применим лемму 1, положив $p = 1$ и $y_1 = \sqrt{n}\bar{x}$.

5.5 Доверительный интервал

Определение 7 Интервал, накрывающий истинное значение параметра распределения с заданной вероятностью P называется $P \cdot 100\%$**-доверительным интервалом.**

В случае нормально распределенной выборки $x_1, \ldots, x_n$ при известной дисперсии σ можно построить доверительный интервал для генерального среднего μ на основании теор. 7,

$$\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \iff \frac{\bar{x}-\mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1)\,.$$

В случае произвольного распределения это выражение можно использовать при достаточно большом объеме выборки вследствие ЦПТ. Зададим уровень значимости α. Для стандартно нормально распределенной случайной величины ξ справедливо

$$P\{|\xi| < z_{1-\frac{\alpha}{2}}\} = 1-\alpha,$$

где через $z_{1-\frac{\alpha}{2}}$ обозначена $(1-\frac{\alpha}{2})$- квантиль нормального распределения $\mathcal{N}(0,1)$. В выражение

$$P\{-z_{1-\frac{\alpha}{2}} < \xi \leq z_{1-\frac{\alpha}{2}}\} = 1-\alpha$$

подставим $\xi = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$,

$$P\{-z_{1-\frac{\alpha}{2}} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}}\} = 1-\alpha$$

и построим неравенство для μ

$$P\left\{\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha\,.$$

Задание для самостоятельной работы. Построить 95%- й доверительный интервал суммарного выигрыша в случае известной дисперсии.

5.6 Статистики Стьюдента и Фишера

Пусть $x_1, \ldots, x_n$ — выборка из нормальной совокупности со средним μ и дисперсией σ^2. Тогда отношение

$$\tau = \frac{\bar{x}-\mu}{\sqrt{m_2}}\sqrt{n-1} \tag{42}$$

имеет распределение Стьюдента с $n-1$ степенью свободы.

Действительно, в (37) полагая $\xi_1 = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ и $\sum\limits_{i=1}^{n} \xi_i^2 = \frac{nm_2}{\sigma^2}$, получим

$$\tau = \frac{\xi_1}{\sqrt{\frac{1}{n-1}\sum\limits_{k=2}^{n}\xi_k^2}} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{1}{n-1}\frac{nm_2}{\sigma^2}}} = \frac{\bar{x}-\mu}{\sqrt{m_2}}\sqrt{n-1}.$$

Так как $m_2 = \frac{n-1}{n}S^2$, $\sqrt{m_2} = \frac{\sqrt{n-1}}{\sqrt{n}}S$,

$$\tau = \frac{\bar{x}-\mu}{S}\sqrt{n} \ \sim \ \mathbf{T(n-1)}. \tag{43}$$

Статистика Фишера для проверки гипотезы о равенстве дисперсий строится с учетом структуры распределения Фишера (40) как функция от двух независимых нормально распределенных выборок: $x_1, \ldots, x_n \sim \mathcal{N}(\mu_1; \sigma)$ и $y_1, \ldots, y_m \sim \mathcal{N}(\mu_2; \sigma)$

$$F = \frac{S_x^2}{S_y^2} \ \sim \ \mathbf{F(n;m)}, \tag{44}$$

где S_x^2 и S_y^2 – несмещенные оценки их дисперсий соответствующих выборок. Действительно, так как $(n-1)S^2/\sigma^2 \ \sim \ \boldsymbol{\chi^2(n-1)}$, то в соответствии с (40)

$$F = \frac{\frac{1}{n-1}(n-1)S_x^2/\sigma^2}{\frac{1}{m-1}(m-1)S_y^2/\sigma^2} = \frac{S_x^2}{S_y^2} \ \sim \ \mathbf{F(n;m)}.$$

Задание для самостоятельной работы. Построить 95%- й доверительный интервал суммарного выигрыша в случае дисперсии, оцениваемой по выборке.

5.7 Критерий Стьюдента для зависимых выборок

Часто встречаются данные, в которых за одними и теми же индивидами наблюдают в разные моменты времени, например,

у $n = 67$ больных с воспалением легких логарифм показателя скорости оседания эритроцитов (СОЭ) в начале заболевания в среднем равен 3.82 со стандартным отклонением 0.13, а при выписке из стационара 3.28 со стандартным отклонением 0.19. Требуется выяснить, можно ли считать уменьшение СОЭ значимым, или это уменьшение можно объснить случайностью.

Итак, имеется парная выборка (x_i, y_i), $i = 1, \ldots, n$, зависимых наблюдений объема n над случайными величинами ξ_1 и ξ_2 с нормальным распределением $\mathcal{N}(\mu_1, \sigma_1)$ и $\mathcal{N}(\mu_2, \sigma_2)$, относительно которых нет предположения о независимости — скорее всего, чем меньше был уровень СОЭ в начале заболевания, тем меньше он при выздоровлении. Рассмотрим разность $\eta = \xi_1 - \xi_2$ с соответствующей выборкой

$$z_i = x_i - y_i.$$

Если справедлива нулевая гипотеза $H_0 : \mu_1 = \mu_2$, то для случайной выборки $z_i = x_i - y_i$ должна быт справедлива гипотеза о равенстве нулю ее математического ожидания. В этом случае для z_i может быть применена статистика (43).

Например, $\bar{z} = 0.54$, $S = 0.19$,

$$t = \frac{\bar{z} - 0}{S}\sqrt{n} = \frac{0.54}{0.19}\sqrt{66} = 23.45,$$

двусторонний доверительный уровень вероятности равен

$$p = P\{|t| > |23.45|\} = \text{СТЮДРАСП}(23.45; 66; 2) = \\ = 10^{-33} < 0.05,$$

поэтому гипотеза $H_0 : \mu_1 = \mu_2$ отвергается, и уменьшение уровня СОЭ в процессе выздоровления нельзя объяснить случайностью.

5.8 Распределение разности выборочных средних

Пусть имеется две независимые выборки $x_1, \ldots, x_{n_1}$ и $y_1, \ldots, y_{n_2}$, о которых известно, что они имеют нормальное распределение со средними μ_1 и μ_2 и одинаковой дисперсией σ^2. Выборочные средние обозначим через $\bar{x}$ и $\bar{y}$, а несмещенные оценки дисперсии через S_1^2 и S_2^2. Тогда статистика

$$T = \frac{(\bar{x} - \bar{y} - (\mu_1 - \mu_2))\sqrt{n_1 + n_2 - 2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}} \sim \mathbf{T}(n_1 + n_2 - 2) \quad (45)$$

имеет распределение Стьюдента с числом степеней свободы, равным $n_1 + n_2 - 2$.

Во-первых, из независимости выборок выборочные средние $\bar{x}$ и $\bar{y}$ также независимы, а по теореме Фишера (т.7) нормально распределены соответственно с параметрами $\mathcal{N}(\mu_1, \sigma/\sqrt{n_1})$ и $\mathcal{N}(\mu_2, \sigma/\sqrt{n_2})$. Следовательно, их разность имеет нормальное распределение с параметрами

$$\bar{x} - \bar{y} \quad \sim \quad \mathcal{N}\left(\mu_1 - \mu_2, \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right),$$

откуда имеем

$$\xi = \frac{\bar{x} - \bar{y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}(0; 1). \quad (46)$$

С другой стороны, также по теореме 7 имеем

$$(n_1 - 1)S_1^2/\sigma^2 \sim \chi^{\mathbf{2}}(\mathbf{n_1 - 1}),$$
$$(n_2 - 1)S_2^2/\sigma^2 \sim \chi^{\mathbf{2}}(\mathbf{n_2 - 1}).$$

Из аддитивности распределения $\chi^{\mathbf{2}}$ их сумма вида

$$\eta = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2} \ \sim \ \boldsymbol{\chi^2(n_1+n_2-2)} \qquad (47)$$

имеет распределение хи-квадрат с числом степеней свободы, равным сумме степеней свободы слагаемых.

Рассмотрим случайную величину вида

$$T = \frac{\xi}{\sqrt{\frac{1}{n_1+n_2-2}\eta}},$$

которая, согласно структуре (37), имеет распределение Стьюдента с числом степеней свободы, равным n_1+n_2-2. Подставив соответствующие ξ из (46) и η из (47), получим вид статистики из (45).

В случае известных дисперсий σ_1^2 и σ_2^2 (возможно неодинаковых), учитывая нормальность средних

$$\bar{x} \sim \mathcal{N}(\mu_1, \sigma_1/\sqrt{n_1}), \quad \bar{y} \sim \mathcal{N}(\mu_2, \sigma_2/\sqrt{n_2}),$$

вычисляем дисперсию разности

$$\mathbf{D}(\bar{x}-\bar{y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

откуда получаем

$$Z = \frac{\bar{x}-\bar{y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \ \sim \ \mathcal{N}(0;1). \qquad (48)$$

В случае неизвестных неодинаковых дисперсий $\sigma_1^2 \neq \sigma_2^2$ статистика T имеет приближенно распределение Стьюдента с нецелым числом степеней свободы.

$$T = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \ \sim \ \mathbf{T}(k), \qquad (49)$$

$$\text{где } k = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}.$$

Заметим, что

$$\mathbf{E}\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}=\sigma^2,$$

так как

$$\mathbf{E}\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2}=n_1+n_2-2$$

числу степеней свободы распределения хи-квадрат.

5.9 Однофакторный дисперсионный анализ

5.9.1 Основные обозначения (л.9)

Пусть имеется выборка наблюдений, которая разбивается на r групп. Каждая группа содержит n_i величин $\mathcal{N}(\mu_i, \sigma)$, где σ не зависит от i, $n_1 + \ldots + n_r = n$. Требуется проверить о равенстве средних $H_0 : \mu_1 = \ldots = \mu_r$.

- Выборка представляет собой вектор размерности n вида

$$x = (x_{11}, \ldots, x_{1n_1}, \ldots, x_{r1}, \ldots, x_{rn_r})^T; \tag{50}$$

- x_{ij} — j-я величина ($j = 1, \ldots, n_i$) в i-й группе ($i = 1, \ldots, r$);
- $\bar{x}_i = \frac{1}{n_i} \sum\limits_{j=1}^{n_i} x_{ij}$ — среднее в i-й группе;
- $\bar{x} = \frac{1}{n} \sum\limits_{i=1}^{r} \sum\limits_{j=1}^{n_i} x_{ij}$ — общее среднее.

5.9.2 Разбиение суммы квадратов отклонений

Покажем, что общий источник вариации Q в виде суммы квадратов отклонений от общего среднего можно разложить на сумму источников вариации, обусловленных различием между группами Q_1 и внутри групп Q_2.

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x})^2 =$$

$$= \underbrace{\sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}_{Q_2} + \underbrace{\sum_{i=1}^{r} \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2}_{Q_1}, \text{ так как}$$

$$\sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(\bar{x}_i - \bar{x}) = \sum_{i=1}^{r} (\bar{x}_i - \bar{x}) \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) = 0\,.$$

Предложение 7 *Ранги квадратичных форм Q, Q_1 и $Q-2$ равны $n-1$, $r-1$ и $n-r$ соответственно.*

Для доказательства введем преобразование $y = Ax$ вектора x из (50) при помощи ортогональной матрицы A, $A^{-1} = A^T$, у которой последняя строка имеет вид

$$\frac{1}{\sqrt{n}} + \ldots + \frac{1}{\sqrt{n}}.$$

Тогда $x = A^{-1}y = A^T y$,

$$\sum_{ij} x_{ij}^2 = x^T x = (A^T y)^T A^T y = y^T A A^T y = y^T y = \sum_{k=1}^{n} y_k^2\,,$$

а квадраточная форма

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2 = \sum_{ij} x_{ij}^2 - n\bar{x}^2 =$$

$$= \sum_{ij} x_{ij}^2 - \left(\frac{x_1}{\sqrt{n}} + \ldots + \frac{x_n}{\sqrt{n}}\right)^2 = \sum_{k=1}^{n} y_k^2 - y_n^2 = \sum_{k=1}^{n-1} y_k^2$$

имеет ранг $n-1$. Источник вариации, обусловленный различием между группами, можно представить в виде

$$Q_1 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{r} L_i^2, \quad \text{где} \quad L_i = \sqrt{n_i}(\bar{x}_i - \bar{x}),$$

$$\sum_{i=1}^{r} L_i\sqrt{n_i} = 0, \quad \text{так как} \quad n\bar{x} = \sum_{i=1}^{r} n_i\bar{x}_i\,.$$

Следовательно, $rank(Q_1) \le r-1$. Источник вариации, обусловленный различием внутри групп, можно представить в виде

$$Q_2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 = \sum_{i,j} L_{ij}^2, \text{ где } L_{ij} = x_{ij} - \bar{x}_i,$$

$$\sum_{j=1}^{n_i} L_{ij} = 0 \;\; \forall i = 1, \ldots, r. \Longrightarrow \; rank(Q_2) \le n - r.$$

Воспользуемся свойством рангов квадратичных форм

$$rank(Q_1 + Q_2) \le rank(Q_1) + rank(Q_2)\,.$$

Следовательно,

$$n - 1 \le (r - 1) + (n - r) = n - 1,$$

и ранги Q_1 и $Q - 2$ в точности равны $r - 1$ и $n - r$.

Воспользуемся теоремой Кохрена: *пусть* Q_i, $i = 1, \ldots, k$, — *неотрицательные квадратичные формы ранга* r_i,

$$\sum_{i=1}^{n} x_i^2 = Q_1 + \ldots + Q_k.$$

Если $\sum\limits_{i=1}^{k} r_i = n$, *то существует ортогональное преобразование* $x = Cy$, *переводящее все* Q_i *в суммы квадратов такого вида:*

$$Q_1 = \sum_{i=1}^{r_1} y_i^2,\ Q_2 = \sum_{i=r_1+1}^{r_1+r_2} y_i^2, \ldots,\ Q_k = \sum_{i=n-r_k+1}^{n} y_i^2,$$

где никакая пара форм не содержит общей переменной y_i.

Отсюда из $Q = Q_1 + Q_2$ при помощи ортогонального преобразования получаем

$$\sum_{i=1}^{n-1} y_i^2 = \sum_{i=1}^{r-1} y_i^2 + \sum_{i=r}^{n-1} y_i^2.$$

Из независимости величин $x_{11}, \ldots, x_{rn_r}$ следует независимость $y_1, \ldots, y_{n-1}$, следовательно, Q_1 и Q_2 независимы.

5.9.3 Статистика критерия Фишера

Предположим, что нулевая гипотеза $H_0 : \mu_1 = \ldots = \mu_r = \mu$ верна. Тогда $x_{ij} = \mu + \xi_{ij}$, где $\xi_{ij} \sim \mathcal{N}(0, 1)$ независимы.

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (\xi_{ij} - \bar{\xi})^2,$$

$$Q_1 = \sum_{i=1}^{r} n_i(\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{r} n_i(\bar{\xi}_i - \bar{\xi})^2,$$
$$Q_2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\xi_{ij} - \bar{\xi}_i)^2.$$

При ортогональном преобразовании получаем, что

$$Q = \sum_{k=1}^{n-1}\eta_k^2, \quad Q_1 = \sum_{k=1}^{r-1}\eta_k^2, \quad Q_2 = \sum_{k=r}^{n-1}\eta_k^2,$$

где $\eta_k \sim \mathcal{N}(0,\sigma)$ независимы. Таким образом, при справедливости нулевой гипотезы $\frac{Q_1}{\sigma^2}$ и $\frac{Q_2}{\sigma^2}$ имеют распределение хи-квадрат с числом степеней свободы соответственно $r-1$ и $n-r$. Отсюда получаем статистику

$$F = \frac{\frac{1}{r-1}Q_1}{\frac{1}{n-r}Q_2} \;\sim\; F(r-1, n-r)\,,$$

которая используется на практике для проверки гипотезы однородности в случае нескольких групп.

5.9.4 Мера изменчивости средних

В случае, когда нулевая гипотеза не верна, рассматриваем $x_{ij} = \mu_i + \xi_{ij}$, где $\xi_{ij} \sim \mathcal{N}(0,\sigma)$, откуда $\bar{x}_i = \mu_i + \bar{\xi}_i$; $\bar{x} = \mu + \bar{\xi}$.

$$(x_{ij} - \bar{x}_i)^2 = (\xi_{ij} - \bar{\xi}_i)^2\,, \tag{51}$$
$$(\bar{x}_i - \bar{x})^2 = (\mu_i + \bar{\xi}_i - \mu - \bar{\xi})^2 =$$
$$= (\bar{\xi}_i - \bar{\xi})^2 + 2(\bar{\xi}_i - \bar{\xi})(\mu_i - \mu) + (\mu_i - \mu)^2\,. \tag{52}$$

Вычислим математические ожидания Q_1 и Q_2. Из (52) получаем

$$\mathbf{E}Q_1 = \mathbf{E}\sum_{i=1}^{r} n_i(\bar{x}_i - \bar{x})^2 =$$
$$= \sum_{i=1}^{r} n_i \left(\mathbf{E}(\bar{\xi}_i - \bar{\xi})^2 + 2\underbrace{\mathbf{E}(\bar{\xi}_i - \bar{\xi})}_{=0}(\mu_i - \mu) + \mathbf{E}(\mu_i - \mu)^2 \right) =$$

$$= \sum_{i=1}^{r} n_i \left(\underbrace{\mathbf{E}\bar{\xi}_i^2}_{\frac{\sigma^2}{n_i}} - \underbrace{\mathbf{E}\bar{\xi}^2}_{\frac{\sigma^2}{n}} \right) + \sum_{i=1}^{r} n_i(\mu_i - \mu)^2 =$$

$$= (r-1)\sigma^2 + \sum_{i=1}^{r} n_i(\mu_i - \mu)^2 .$$

Поясним то, как было получено первое слагаемое.

$$\mathbf{E}(\bar{\xi}_i - \bar{\xi})^2 = \mathbf{E}(\bar{\xi}_i^2 - 2\bar{\xi}_i\bar{\xi} + \bar{\xi}^2),$$

$$\mathbf{E}\bar{\xi}_i^2 = \frac{\sigma^2}{n_i}, \quad \mathbf{E}\bar{\xi}^2 = \frac{\sigma^2}{n} .$$

$$\mathbf{E}\bar{\xi}_i\bar{\xi} = \mathbf{E}\bar{\xi}_i \left(\frac{1}{n} \sum_{j=1}^{r} n_j \bar{\xi}_j \right) = \frac{1}{n} \sum_{j=1}^{r} n_j \mathbf{E}\bar{\xi}_i\bar{\xi}_j = \frac{n_i \mathbf{E}\bar{\xi}_i^2}{n} = \frac{\sigma^2}{n} .$$

Аналогично математическое ожидание Q_2 имеет вид

$$\mathbf{E}Q_2 = \sum_{i,j} \mathbf{E}(x_{ij} - \bar{x}_i)^2 = \sum_{i,j} \mathbf{E}(\xi_{ij} - \bar{\xi}_i)^2 =$$

$$= \sum_{i,j} [\underbrace{\mathbf{E}\xi_{ij}^2}_{=\sigma^2} - \underbrace{\mathbf{E}\bar{\xi}_i^2}_{=\frac{\sigma^2}{n_i}}] = \left(n\sigma^2 - \sum_{i=1}^{r} n_i \frac{\sigma^2}{n_i} \right) = (n-r)\sigma^2 .$$

Таким образом, для $s_1^2 = \frac{Q_1}{r-1}$ и $s_2^2 = \frac{Q_2}{n-r}$ получаем математические ожидания

$$\mathbf{E}s_1^2 = \sigma^2 + \frac{1}{r-1} \sum_{i=1}^{r} n_i(\mu_i - \mu)^2, \qquad \mathbf{E}s_2^2 = \sigma^2 .$$

Выражение $\frac{1}{n} \sum\limits_{i=1}^{r} n_i(\mu_i - \mu)^2$ может быть использовано в качестве *меры изменчивости средних.* Его оценой является статистика

$$\frac{r-1}{n}(s_1^2 - s_2^2),$$

так как ее математическое ожидание равно

$$\frac{r-1}{n}\mathbf{E}(s_1^2 - s_2^2) = \frac{r-1}{n} \cdot \frac{1}{r-1} \sum_{i=1}^{r} n_i(\mu_i - \mu)^2 .$$

5.9.5 Наведение контрастов

Для проверки $H_0 : \mu_i = \mu_k$ значимости отклонений внутригрупповых средних используем отношение Стьюдента

$$t = \frac{\xi_0}{\sqrt{\eta}} = \frac{\xi_0}{\sqrt{\frac{1}{M}\zeta}},$$

где $\xi_0 \sim \mathcal{N}(0,1),\ \zeta \sim \chi^2(M)$. Для выражения ξ_0 рассмотрим разность $\bar{x}_i - \bar{x}_k$ с характеристиками:

$$\begin{aligned}
\mathbf{E}(\bar{x}_i - \bar{x}_k) &= \mu_i - \mu_k, \\
\mathbf{D}(\bar{x}_i - \bar{x}_k) &= \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_k}, \\
\xi_0 &= \frac{(\bar{x}_i - \bar{x}_k) - (\mu_i - \mu_k)}{\sigma\sqrt{\frac{1}{n_i} + \frac{1}{n_k}}}.
\end{aligned}$$

Так как $Q_1 = \sum_i n_i(\bar{x}_i - \bar{x})^2$ не зависит от Q_2, при ортогональном преобразовании переменных x_{ij} в $y_k,\ k = 1, \ldots, n$, разности вида

$$\bar{x}_i - \bar{x}_k = (\bar{x}_i - \bar{x}) - (\bar{x}_k - \bar{x})$$

выражаются через линейные комбинации переменных $y_1, \ldots, y_r$ и не зависят от Q_2, выражаемой через переменные $y_{r+1}, \ldots, y_n$. Следовательно, в отношении Стьюдента можно использовать выражение $\eta = \frac{s_2^2}{\sigma^2}$, где

$$\begin{aligned}
s_2^2 = \frac{Q_2}{(n-r)}, \quad \frac{Q_2}{\sigma^2} &\sim \chi^{\mathbf{2}}(n-r), \\
t = \frac{(\bar{x}_i - \bar{x}_k) - (\mu_i - \mu_k)}{Q_2\sqrt{\frac{1}{n_i} + \frac{1}{n_k}}}\sqrt{n-r} \ &\sim \ T(n-r)\,.
\end{aligned}$$

6 Случай двух измерений

6.1 Регрессия

Пусть случайные величины ξ и η имеют совместное распределение непрерывного типа с плотностью $f(x,y)$, для любого x частная плотность имеет вид

$$f_1(x) = \int\limits_{-\infty}^{+\infty} f(x,y)dy > 0\,,$$

а условная плотность вероятности

$$f(y|x) = \frac{f(x,y)}{f_1(x)}$$

есть распределение массы в бесконечно узкой вертикальной полосе, проходящей через точку $\xi = x$. *Кривую регрессии для среднего величины* η определяют как траекторию точки (x,y) при переменном x, где

$$y = \mathbf{E}(\eta|\xi = x)\,. \tag{53}$$

Кривые регрессии обладают свойством минимальности. Среди всех возможных функций $g(\xi)$ минимум выражения

$$\mathbf{E}(\eta - g(\xi))^2 = \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty}(y-g(x))^2 f(x,y)dxdy = \tag{54}$$
$$= \int\limits_{-\infty}^{+\infty} f_1(x)dx \int\limits_{-\infty}^{+\infty}(y-g(x))^2 f(y|x)dy$$

достигается при функции $g(\xi) = \mathbf{E}(\eta|\xi)$, изображаемой в виде кривой (53). Если кривая регрессии есть прямая, то имеет место случай *линейной регрессии.*

Если в экстремальной задаче (54) ограничиться функциями, принадлежащему определенному классу, например, линейными функциями, полиномами и т.д., то кривые, полученные этим путем, называются *кривыми средней квадратической регрессии.*

6.2 Оценка параметров линейной средней квадратической регрессии

Рассмотрим двумерную выборку наблюдений (x_i, y_i) над величинами ξ и η, $i = 1, \ldots, n$. Из всего семейства прямых вида $g(x) = \alpha + \beta x$ выбирается „наилучшая“ в том смысле, что квадратичная форма

$$Q = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

принимает наименьшее значение. В регрессионной модели

$$y_i = \alpha + \beta x_i + \delta_i \tag{55}$$

считаем x_i детерминированными, а δ_i – центрированными, некоррелированными случайными величинами с дисперсией σ^2,

$$\mathbf{E}\delta_i = 0, \quad \mathbf{D}\delta_i = \sigma^2, \quad \mathbf{E}\delta_i\delta_j = 0. \tag{56}$$

Случайные величины δ_i называют ошибками, так как они показывают, насколько велико отклонение наблюдений y_i от линии регрессии $y = \alpha + \beta x$.

Для оценки параметров α и β следует решить систему нормальных уравнений, полученную в результате дифференцирования квадратичной формы Q по параметрам и приравнивания полученных выражений нулю:

$$\begin{cases} -\frac{1}{2}\frac{\partial Q}{\partial \alpha} = \sum\limits_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0 \\ -\frac{1}{2}\frac{\partial Q}{\partial \beta} = \sum\limits_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i = 0 \end{cases} \iff$$

$$\iff \begin{cases} \sum\limits_{i=1}^{n} y_i - \alpha n - \beta \sum\limits_{i=1}^{n} x_i = 0 \\ \sum\limits_{i=1}^{n} y_i x_i - \alpha \sum\limits_{i=1}^{n} x_i - \beta \sum\limits_{i=1}^{n} x_i^2 = 0 \end{cases}$$

Учитывая то, что $\sum\limits_{i=1}^{n} x_i = n\bar{x}$ и $\sum\limits_{i=1}^{n} y_i = n\bar{y}$, получаем оценки параметров линейной регрессии:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \hat{\beta} = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2}. \tag{57}$$

Не умаляя общности, будем считать

$$\sum_{i=1}^{n} x_i = 0. \tag{58}$$

При этом оценки параметров линейной регрессии упрощаются и имеют вид

$$\hat{\alpha} = \bar{y}, \quad \hat{\beta} = \frac{\sum\limits_{i=1}^{n} x_i y_i}{\sum\limits_{i=1}^{n} x_i^2}. \tag{59}$$

6.2.1 Распределение $\hat{\alpha}$ и $\hat{\beta}$ в случае известной дисперсии ошибок σ^2

Предложение 8 *Пусть ошибки регрессии распределены нормально, то есть* $\delta_i \sim \mathcal{N}(0;\sigma)$, $\sum\limits_{i=1}^{n} x_i = 0$. *Обозначим через* $[x,x]$ *скалярное произведение* $[x,x] = \sum\limits_{i=1}^{n} x_i^2$. *Тогда для оценок (59) справедливо*

$$\begin{aligned} &1) \ \ \hat{\alpha} \sim \mathcal{N}\left(\alpha, \frac{\sigma}{\sqrt{n}}\right), \\ &2) \ \ \hat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma}{\sqrt{[x,x]}}\right). \end{aligned} \tag{60}$$

Действительно,

$$\hat{\alpha} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i + \delta_i) = \\ = \alpha + \frac{\beta}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n}\delta_i.$$

Так как $\sum\limits_{i=1}^{n} x_i = 0$ в силу (58), то получаем

$$\hat{\alpha} = \alpha + \bar{\delta},$$

где $\bar{\delta} = \frac{1}{n}\sum\limits_{i=1}^{n}\delta_i$. Согласно (56), получаем $\mathbf{E}\hat{\alpha} = \alpha$, $\mathbf{D}\hat{\alpha} = \sigma^2/n$, $\hat{\alpha} \sim \mathbf{N}(\alpha, \sigma/\sqrt{n})$.

Далее, выразим оценку $\hat{\beta}$ через ошибки δ_i. Подставим (55) в выражение (56)оценки $\hat{\beta}$.

$$\hat{\beta} = \frac{\sum\limits_i y_i x_i}{[x,x]} = \frac{\sum\limits_i (\alpha + \beta x_i + \delta_i)x_i}{[x,x]} = \frac{\alpha\sum\limits_i x_i + \beta\sum\limits_i x_i^2 + \sum\limits_i \delta_i x_i}{[x,x]}.$$

Из $\sum\limits_{i=1}^{n} x_i = 0$ и $[x,x] = \sum\limits_{i=1}^{n} x_i^2$ получаем

$$\hat{\beta} = \beta + \frac{1}{[x,x]}\sum_i \delta_i x_i, \qquad \mathbf{E}\hat{\beta} = \beta,$$

$$\mathbf{D}\hat{\beta} = \mathbf{D}\left(\beta + \frac{1}{[x,x]}\sum_i \delta_i x_i\right) = \frac{1}{[x,x]^2}\sum_i x_i^2 \mathbf{D}\delta_i = \frac{1}{[x,x]}\sigma^2.$$

Заметим, что

$$\mathbf{E}(\hat{\alpha} - \alpha)(\hat{\beta} - \beta) = \mathbf{E}\left(\frac{1}{n}\sum_i \delta_i \cdot \frac{1}{[x,x]}\sum_j x_j\delta_j\right) = \\ = \frac{1}{n[x,x]}\sum_i x_i \mathbf{E}\delta_i^2 = \frac{\sigma^2}{n[x,x]}\sum_i x_i = 0\,.$$

6.2.2 Распределение $\hat{\alpha}$ и $\hat{\beta}$ в случае неизвестной дисперсии ошибок σ^2

Обозначим через $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ оценку линии регрессии и через

$$Q_E = \sum_i (y_i - \hat{y}_i)^2 \tag{61}$$

сумму квадратов отклонений наблюдаемых значаний y_i от линии регрессии $\hat{y}_i$.

Предложение 9 *Случайная величина* $\frac{Q_E}{\sigma^2}$ *имеет распределение* $\chi^{\mathbf{2}}(\mathbf{n-2})$.

Обозначим через $d_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$. Используя то, что $\sum_i x_i = 0$, $\hat{\alpha} = \bar{y}$, $\hat{\beta} = \frac{\sum\limits_{i=1}^{n} x_i y_i}{\sum\limits_{i=1}^{n} x_i^2}$, убеждаемся, что

$$\sum_i d_i = \sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0,$$
$$\sum_i x_i d_i = \sum_i x_i (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0\,.$$

На основании этого

$$\sum_i \delta_i^2 = \sum_i (y_i - \alpha - \beta x_i)^2 =$$
$$= \sum_i ((y_i - \hat{\alpha} - \hat{\beta} x_i) + (\hat{\alpha} - \alpha) + (\hat{\beta} - \beta) x_i)^2 =$$
$$= Q_E + n(\hat{\alpha} - \alpha)^2 + (\hat{\beta} - \beta)^2 [x, x].$$

При этом перекрестные суммы оказываются равными нулю:

$$\sum_i d_i (\hat{\alpha} - \alpha) = 0,$$
$$\sum_i d_i x_i (\hat{\beta} - \beta) = 0,$$
$$\sum_i (\hat{\alpha} - \alpha)(\hat{\beta} - \beta) x_i = 0\,.$$

Согласно лемме 1, в которой в

$$x_i = \delta_i,$$

$$y_1 = \sqrt{n}(\hat{\alpha} - \alpha) = \sqrt{n}\bar{\delta} = \frac{1}{\sqrt{n}}(\delta_1 + \ldots + \delta_n),$$

$$y_2 = \sqrt{[x,x]}(\hat{\beta} - \beta) = \frac{1}{\sqrt{[x,x]}} \sum_{i=1}^{n} x_i \delta_i,$$

квадратичная форма Q_E имеет ранг $n-2$, и $\frac{Q_E}{\sigma^2}$ имеет распределение хи-квадрат с $n-2$ степенями свободы и является несмещенной оценкой дисперсии ошибки

$$\hat{\sigma}^2 = \frac{Q_E}{n-2} = S^2. \tag{62}$$

Используя

$$\mathbf{E}\delta_i^2 = \sigma^2, \quad \mathbf{E}(\hat{\alpha} - \alpha)^2 = \frac{\sigma^2}{n}, \quad \mathbf{E}(\hat{\beta} - \beta)^2 = \frac{\sigma^2}{[x,x]},$$

$$\mathbf{E}(\hat{\alpha} - \alpha)\delta_i = \mathbf{E}\frac{1}{n}\sum_j \delta_j \delta_i = \frac{\sigma^2}{n},$$

$$\mathbf{E}(\hat{\beta} - \beta)\delta_i x_i = \mathbf{E}\frac{1}{[x,x]}\sum_j x_j \delta_j \delta_i x_i = \frac{x_i^2 \sigma^2}{[x,x]},$$

$$\mathbf{E}(\hat{\beta} - \beta)(\hat{\alpha} - \alpha) = 0,$$

покажем, что $\mathbf{E}Q_E = (n-2)\sigma^2$. Действительно,

$$\begin{aligned}
\mathbf{E}\sum_i (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 &= \mathbf{E}\sum_i (\alpha + \beta x_i + \delta_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \\
&= \mathbf{E}\sum_i (\delta_i - (\hat{\alpha} - \alpha) - (\hat{\beta} - \beta)x_i)^2 = \\
&= \sum_i \mathbf{E}\delta_i^2 + n\mathbf{E}(\hat{\alpha} - \alpha)^2 + \mathbf{E}(\hat{\beta} - \beta)^2 \sum_i x_i^2 - \\
&\quad -2\mathbf{E}(\hat{\alpha} - \alpha)\sum_i \delta_i - 2\mathbf{E}(\hat{\beta} - \beta)\sum_i \delta_i x_i + \\
&\quad +2\mathbf{E}(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)\sum_i x_i = \\
&= n\sigma^2 + \sigma^2 + \sigma^2 - 2\sigma^2 - 2\sigma^2 + 0 = (n-2)\sigma^2.
\end{aligned}$$

Предложение 10 *Пусть* $\delta_i \sim \mathbf{N}(0;\sigma)$, *где* σ *неизвестна,* $S^2 = \frac{Q_E}{n-2}$, $S_\alpha^2 = S^2/n$, $S_\beta^2 = S^2/[x,x]$. *Тогда*

$$T' = \frac{\hat{\alpha}-\alpha}{S_\alpha} \sim \mathbf{T}(n-2), \tag{63}$$

$$T'' = \frac{\hat{\beta}-\beta}{S_\beta} \sim \mathbf{T}(n-2), \tag{64}$$

В случае нецентрированных данных, когда $\sum\limits_{i=1}^{n} x_i \neq 0$, *имеем* $[x,x] = \sum\limits_{i=1}^{n}(x_i - \bar{x})^2$,

$$S_\alpha^2 = \frac{S^2 \sum\limits_{i=1}^{n} x_i^2}{n[x,x]}. \tag{65}$$

Доказательство. 1) Из предл. 8 имеем

$$U_1 = \frac{\hat{\alpha}-\alpha}{\sigma/\sqrt{n}} \sim \mathbf{N}(0;1).$$

Из предл. 9 и выражения (62) получаем, что

$$\chi^2 = (n-2)S^2/\sigma^2 \sim \boldsymbol{\chi^2}(n-2).$$

Исходя из этого построим случайную величину T', имеющую распределение Стьюдента:

$$T' = \frac{U_1}{\sqrt{\frac{1}{n-2}\chi^2}} = \frac{(\hat{\alpha}-\alpha)\sqrt{n}}{\sigma \cdot \sqrt{\frac{1}{n-2}(n-2)S^2/\sigma^2}} = \\ = \frac{\hat{\alpha}-\alpha}{S}\sqrt{n} \sim \mathbf{T}(n-2).$$

2) Согласно предл. 8 имеем

$$U_2 = \frac{\hat{\beta}-\beta}{\sigma/\sqrt{[x,x]}} \sim \mathbf{N}(0;1).$$

Аналогично построим случайную величину T'', имеющую распределение Стьюдента.

$$T'' = \frac{(\hat{\beta} - \beta)\sqrt{[x,x]}}{\sigma \cdot \sqrt{\frac{1}{n-2}(n-2)S^2/\sigma^2}} =$$
$$= \frac{\hat{\beta} - \beta}{S}\sqrt{[x,x]} \ \sim \ \mathbf{T}(n-2).$$

В случае $\bar{x} \neq 0$ имеем $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$, $[x,x] = \sum\limits_{i=1}^{n}(x_i - \bar{x})^2$. Дисперсия оценки параметра α имеет вид

$$\mathbf{D}\hat{\alpha} = \mathbf{D}(\bar{y} - \hat{\beta}\bar{x}) = \mathbf{D}\bar{y} + \mathbf{D}(\hat{\beta}\bar{x}) =$$
$$= \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{[x,x]} =$$
$$= \sigma^2\frac{\sum x_i^2 - n\bar{x}^2 + n\bar{x}^2}{n[x,x]} = \frac{\sigma^2\sum x_i^2}{n[x,x]}.$$

Таким образом, по выборке дисперсии коэффициентов регрессии могут быть оценены следующим образом:

$$\mathbf{D}\hat{\alpha} = S_\alpha^2 = \frac{Q_E\sum\limits_{i=1}^{n} x_i^2}{n(n-2)\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}, \tag{66}$$

$$\mathbf{D}\hat{\beta} = S_\beta^2 = \frac{Q_E}{(n-2)\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}. \tag{67}$$

S_α и S_β называются стандартными ошибками коэффициентов регрессии.

6.3 Коэффициенты корреляции и детерминации

Коэффициент детерминации вводится как

$$R^2 = 1 - \frac{Q_E}{Q}, \tag{68}$$

где $Q = \sum\limits_{i=1}^{n}(y_i - \bar{y})^2$ – общая сумма квадратов отклонений от выборочного среднего.

Предложение 11 *Обозначим через Q_R сумму квадратов отклонений линии регрессии от выборочного среднего:*

$$Q_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

Тогда справедливо тождество

$$Q = Q_R + Q_E \tag{69}$$

Действительно, рассмотрим сумму квадратов

$$Q = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 =$$
$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) =$$
$$= Q_E + Q_R.$$

Дополнительное слагаемое равно нулю. Это нетрудно показать. Так как $\hat{y}_i - \bar{y} = \hat{\beta}(x_i - \bar{x})$, то

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n}\hat{\beta}(y_i - \hat{y}_i)(x_i - \bar{x}) =$$
$$= \hat{\beta}\sum_{i=1}^{n}(y_i - \bar{y} - \hat{\beta}(x_i - \bar{x}))(x_i - \bar{x}) =$$
$$= \hat{\beta}(\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x})^2) = 0,$$

поскольку для оценки параметра $\hat{\beta}$ справедливо выражение

$$\hat{\beta} = \frac{\sum\limits_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum\limits_{i=1}^{n}x_i y_i - n\bar{x}\bar{y}}{\sum\limits_{i=1}^{n}x_i^2 - n\bar{x}^2}. \tag{70}$$

Выборочный коэффициент корреляции, полученный по двумерной выборке $(x_1, y_1), \ldots, (x_n, y_n)$, вычисляется по формуле:

$$\hat{\varrho} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2}} \tag{71}$$

Предложение 12 *Коэффициент детерминации равен квадрату коэффициента корреляции:* $R^2 = \hat{\varrho}^2$

Действительно,

$$R^2 = 1 - \frac{Q_E}{Q} = \frac{Q_R}{Q} = \frac{\sum\limits_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\hat{\beta}\sum\limits_{i=1}^{n}(\hat{x}_i - \bar{x})^2}{\sum\limits_{i=1}^{n}(\hat{y}_i - \bar{y})^2}.$$

Подставим (70) вместо $\hat{\beta}$ и сократим на $\sum\limits_{i=1}^{n}(x_i - \bar{x})^2$,

$$R^2 = \frac{\left(\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\left(\sum\limits_{i=1}^{n}(x_i - \bar{x})^2\right)^2}\frac{\sum\limits_{i=1}^{n}(\hat{x}_i - \bar{x})^2}{\sum\limits_{i=1}^{n}(\hat{y}_i - \bar{y})^2}.$$

6.4 Проверка гипотез в корреляционном анализе

Проверка значимости коэффициента детерминации, то есть гипотезы о том, что $H_0 : R^2 = 0$, эквивалентна проверке гипотез о равенстве нулю коэффициента корреляции $H_0 : \varrho = 0$ и коэффициента регрессии $H_0 : \beta = 0$. Все эти три гипотезы проверяются при помощи статистики Фишера вида

$$F = \frac{Q_R}{Q_E/(n-2)}, \tag{72}$$

которая в случае справедливости вышеуказанных нулевых гипотез H_0 имеет распределение Фишера с 1 и $n-2$ степенями свободы. Из предл.9 следует, что Q_R и Q_E независимы, $Q_E \sim \chi^{\mathbf{2}}(n-2)$,

$$Q_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}^2 \sum_{i=1}^{n}(\hat{x}_i - \bar{x})^2 = \hat{\beta}^2[x,x].$$

Если $\beta = 0$, тогда, согласно (60), $\hat{\beta} \sim \mathbf{N}(0; \sigma/\sqrt{[x,x]})$, поэтому $\xi = \hat{\beta}\sqrt{[x,x]}/\sigma \sim \mathbf{N}(0;1)$, $Q_R/\sigma^2 = \xi^2 \sim \chi^{\mathbf{2}}(1)$.

Статистика $T = \sqrt{F}$ имеет распределение Стьюдента с $n-2$ степенями свободы. Она выражается через коэффициент детерминации и, соответственно, через коэффициент корреляции.

$$T = \sqrt{F} = \sqrt{\frac{Q_R}{Q_E}}\sqrt{n-2} = \frac{\sqrt{Q_R}}{\sqrt{Q - Q_R}}\sqrt{n-2} = \\ = \frac{\sqrt{Q_R/Q}}{\sqrt{1 - Q_R/Q}}\sqrt{n-2}.$$

Таким образом получаем статистику для проверки значимости коэффициента корреляции

$$T = \frac{\hat{\varrho}}{\sqrt{1-\hat{\varrho}^2}}\sqrt{n-2} \ \sim \ \mathbf{T}(n-2), \tag{73}$$

которая в случае справедливости нулевой гипотезы $H_0 : \varrho = 0$ имеет распределение Стьюдента с $n-2$ степенями свободы.

6.5 Доверительный интервал для линии регрессии

Уравнение линии регрессии $\tilde{y} = \hat{\alpha} + \hat{\beta}x$ зависит от случайных величин $\hat{\alpha}$, $\hat{\beta}$, $\mathbf{D}\tilde{y} = \mathbf{D}\hat{\alpha} + x^2\mathbf{D}\hat{\beta}$,

$$\mathbf{D}\hat{\alpha} = \frac{\sigma^2 \sum_i x_i^2}{n\sum_i (x_i - \bar{x})^2}, \quad \mathbf{D}\hat{\beta} = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2},$$
$$\mathbf{D}\tilde{y} = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\left(\frac{1}{n}\sum_i x_i^2 + x^2\right)$$

Для каждого значения x построим статистику Стьюдента

$$t = \frac{\tilde{y} - (\alpha + \beta x)}{\sqrt{\mathbf{D}\tilde{y}}\sqrt{\frac{Q_e}{(n-2)\sigma^2}}} = \frac{\tilde{y} - (\alpha + \beta x)}{S}, \quad S^2 = \frac{Q_E\left(\frac{1}{n}\sum_i x_i^2 + x^2\right)}{(n-2)\sum_i (x_i - \bar{x})^2}$$

Доверительный интервал

$$\alpha + \beta x \pm t_{1-\alpha/2}^{(n-2)} S$$

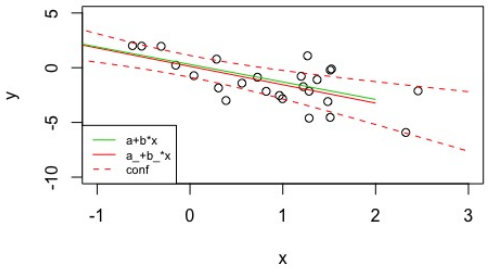

Рис. 4: Линейная регрессия с доверительным интервалом.

7 Непараметрические коэффициенты корреляции

7.1 Ранговая корреляция Спирмена

Довольно часто обычный коэффициент корреляции не вполне отражает структуру взаимосвязи между признаками. Поэтому проверку гипотезы независимости количестенных и ранговых (тех, которые можно упорядочить) признаков рекомендуется всегда дублировать использованием непараметрических методов. Чаще всего используется более доступный в смысле вычислений ранговый коэффициент корреляции Спирмена.

Напомним, что рангом называется порядковый номер элемента в упорядоченной выборке. Пусть $x'_1, \ldots, x'_n$ и $y'_1, \ldots, y'_n$ – ранги индивидов по признакам X и Y, $\xi_i = k_i - l_i$, где

- k_i — количество индивидов, которых превосходит x_i,
- l_i — количество индивидов, которые превосходят x_i.

Аналогично построим η_i. Формально эти характеристики вводятся как

$$\xi_i = (x'_i - 1) - (n - x'_i) = 2\left(x'_i - \frac{n+1}{2}\right),$$
$$\eta_i = (y'_i - 1) - (n - y'_i) = 2\left(y'_i - \frac{n+1}{2}\right).$$

Если признаки взаимосвязаны, то есть большему значению X соответствует большее значение Y, меньшему меньшее и так далее, то ξ_i и η_i совпадают, и характеристика

$$Q_{xy} = \sum_{i=1}^{n} \xi_i \eta_i$$

принимает наибольшее значение, равное

$$Q = Q_{xx} = Q_{yy} = \sum_{i=1}^{n} \xi_i^2 = \sum_{i=1}^{n} \eta_i^2.$$

Если взаимосвязь признаков обратная, то есть меньшему значению X соответствует большее значение признака Y и так далее, то $\xi_i = -\eta_i$, и $Q_{xy} = -Q$. Таким образом можно ввести характеристику

$$R_S = \frac{Q_{xy}}{Q}, \tag{74}$$

которая называется *коэффициентом корреляции Спирмена*. Если между признаками существует прямая зависимость

$$\begin{array}{c} X \\ Y \end{array} \begin{pmatrix} 1 & 2 & \dots & n-1 & n \\ 1 & 2 & \dots & n-1 & n \end{pmatrix},$$

то коэффиент Спирмена равен единице, если обратная

$$\begin{array}{c} X \\ Y \end{array} \begin{pmatrix} 1 & 2 & \dots & n-1 & n \\ n & n-1 & \dots & 2 & 1 \end{pmatrix},$$

то $R_S = -1$.

Можно вычислить непосредственно значение Q.

$$\begin{aligned} Q = \sum_{i=1}^{n} \xi_i^2 = 4 \sum_{i=1}^{n} \left(x_i' - \frac{n+1}{2} \right)^2 = \\ = 4 \left(\sum_{i=1}^{n} (x_i')^2 - 2 \sum_{i=1}^{n} x_i' \frac{n+1}{2} + \frac{n(n+1)^2}{4} \right) = \\ = \frac{4n(n+1)(2n+1)}{6} - 2n(n+1)^2 + n(n+1)^2 = \\ = \frac{n(n+1)}{3}(2(2n+1) - 3(n+1)) = \frac{n(n+1)(n-1)}{3}. \end{aligned}$$

Распределение R_S в случае независимых признаков.

$$\begin{aligned} Q^2 R^2 = Q_{xy}^2 = (\sum \xi_i \eta_i)(\sum \xi_k \eta_k) = \sum_i \sum_k \xi_i \xi_k \eta_i \eta_k, \\ Q^2 \sigma^2 = Q^2 \mathbf{E} R^2 = \sum_i \sum_k \mathbf{E} \xi_i \xi_k \mathbf{E} \eta_i \eta_k = \\ = n \mathbf{E} \xi_1^2 \mathbf{E} \eta_1^2 + n(n-1) \mathbf{E} \xi_1 \xi_2 \mathbf{E} \eta_1 \eta_2. \end{aligned}$$

В силу равенства $\sum \xi_i = 0 = \mathbf{E}\xi_1 \sum \xi_i = \mathbf{E}\xi_1^2 + (n-1)\mathbf{E}\xi_1\xi_i$

$$\mathbf{E}\xi_1\xi_i = -\frac{1}{n-1}\mathbf{E}\xi_1^2, \quad \mathbf{E}\eta_1\eta_i = -\frac{1}{n-1}\mathbf{E}\eta_1^2,$$

$$\sum_i \xi_i^2 = Q, \Longrightarrow n\mathbf{E}\xi_1^2 = Q, \Longrightarrow \mathbf{E}\xi_1^2 = \mathbf{E}\eta_1^2 = \frac{Q}{n},$$

$$Q\sigma^2 = \frac{Q^2}{n} + \frac{Q^2}{n(n-1)} = \frac{Q^2}{n-1}, \Longrightarrow \sigma^2 = \frac{1}{n-1}\,.$$

Поскольку $\frac{n+1}{2}$ представляет собой среднее арифметическое элементов последовательности $1, 2, \ldots, n$, то коэффициент корреляции Спирмена есть не что иное, как коэффицент корреляции между рангами.

Значимость коэффициента Спирмена, как и выборочного коэффициента корреляции, проверяется при помощи статистики (73) вида

$$T = \frac{R_S}{\sqrt{1 - R_S^2}}\sqrt{n-2},$$

которая при отсутствии корреляции между рангами имеет распределение Стьюдента с $n-2$ степенями свободы.

ПРИМЕР. Исследуем зависимость между смертностью от несчастных случаев и других внешних воздействий и объемом валового внутреннего продукта (ВВП) в 1997 году в 21 стране по данным, представленным в таблице 2. Эти страны в силу определенных исторических событий в конце двадцатого века делятся на две группы: Россия, Болгария, Венгрия, Польша, Румыния образуют одну группу, остальные страны другую. Во второй группе наблюдается отрицательная корреляция между объемом ВВП и смертностью от несчастных случаев. Коэффициент корреляции равен $R = -0.46,\ P = 0.07$. Двумерная диаграмма, изображенная на рисунке 5, свидетельствует о неоднородности данных. Ранговый коэффициент корреляции Спирмена между объемом ВВП в 1997 году и смертностью от несчастных случаев во второй группе стран равен $R_S = 0.03$, $P = 0.91$. Поскольку корреляция не значима, то наблюдаемое уменьшение смертности при увеличении объема ВВП является не систематическим, а случайным за счет одной страны (Мексики).

страна	группа	см-ть от несч.сл.	ВВП97	ест.пр.97
Россия	1	187,4	24,2	-5,2
Австрал	2	39,4	76,4	6,7
Австрия	2	54,3	79,5	0,5
Бельгия	2	56,3	79,8	1,2
Болгария	1	66,8	17,9	-7
Англия	2	28,5	70,2	1,6
Венгрия	1	102,4	33,5	-3,8
Германия	2	41,2	76,6	-0,8
Дания	2	55,9	83,2	1,5
Италия	2	40,8	72,8	-0,4
Канада	2	43,6	81,7	4,9
Мексика	2	81,8	25,8	23,4
Нидер-ды	2	29,8	75,6	3,5
Норвегия	2	42,4	91,6	3,5
Польша	1	77	24,7	0,9
Румыния	1	80,8	23,7	-1,9
США	0	55,5	100	5,7
Фин-ия	2	79,1	68,5	1,8
Франция	2	64,5	73,3	3,3
Швеция	2	40,3	71,6	-0,4
Япония	2	43,3	86,2	2,1

Таблица 2: Экономико-демографические данные 1997г.

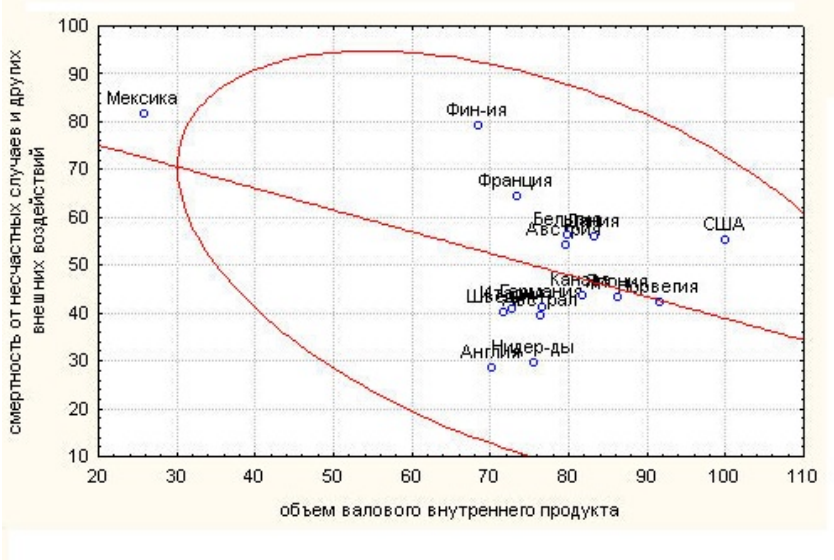

Рис. 5: Наблюдаемая отрицательная зависимость между объемом валового внутреннего продукта и смертностью от несчастных случаев во второй группе стран, $R = -0.46$, $P = 0.07$, но не подтверждаемая при помощи коэффициента корреляции Спирмена, $R_S = 0.03$, $P = 0.91$.

7.1.1 Коэффициент корреляция Кендалла

Рассмотрим две выборки $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$ наблюдений двух признаков, зависимость между которыми нужно как-то измерить. Представим структуру отношений порядков внутри этих выборок в виде элементов матриц порядка n

$$x_{ik} = \begin{cases} 1, & x_i < x_k, \\ 0, & x_i = x_k, \\ -1, & x_i > x_k \end{cases} \quad \text{и} \quad y_{ik} = \begin{cases} 1, & y_i < y_k, \\ 0, & y_i = y_k, \\ -1, & y_i > y_k \end{cases}$$

и введем показатель синхронизации

$$S = \sum_{i=1}^{n} \sum_{k=i+1}^{n} x_{ik} y_{ik} \,.$$

Поскольку число всевозможных пар равно $N = C_n^2 = \frac{n(n-1)}{2}$, то доля синхронизированных пар, называемая *коэффициентом корреляции Кендалла*, равна $T = \frac{S}{N}$.

Например, имееются пять индивидов, которые упорядочены по признаку X (рост), а по признаку Y (вес) выставлены ранги,

$$\begin{array}{c} X \\ Y \end{array} \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 1 & 3 & 2 \end{pmatrix}.$$

Составим таблицу для вычисления синхронизации S.

ik	12	13	14	15	23	24	25	34	35	45
x_{ik}	1	1	1	1	1	1	1	1	1	1
y_{ik}	-1	-1	-1	-1	-1	-1	-1	1	1	-1
$x_{ik}y_{ik}$	-1	-1	-1	-1	-1	-1	-1	1	1	-1

Если обозначить через $S_+ = 2$ и $S_- = 8$ количество 1 и -1 соответственно, тогда показатель синхронизации можно выразить через их разность,

$$S = 1 \cdot S_+ + (-1) \cdot S_- = S_+ - S_-.$$

Можно заметить, что S_+ равен количеству элементов выборки Y, больших y_m по всем m при упорядочивании их по признаку X. Учитывая то, что $S_+ + S_- = N$, получаем

$$S = S_+ - S_- = S_+ - (N - S_+) = 2S_+ - N,$$
$$T = \frac{S}{N} = \frac{2S_+ - N}{N} = \frac{2S_+}{N} - 1\,.$$

Для проверки значимости отклонения от нуля коэффициента корреляции Кендалла используется то, что при независимости признаков T имеет нормальное распределение с нулевым средним и дисперсией

$$\sigma^2 = \frac{2(2n+5)}{9n(n-1)}\,.$$

8 Непараметрические критерии однородности

8.1 Критерии для двух независимых выборок

8.1.1 Критерий W-Вилкоксона (л.10)

Критерий W-Вилкоксона используется для проверки гипотезы однородности двух независимых выборок объема n_1 и n_2, утверждающую, что выборки получены из одной генеральной совокупности.

Для построения статистики критерия выборки $x_1, \ldots, x_{n_1}$ и $y_1, \ldots, y_{n_2}$ объединяют в одну совокупность и упорядочивают ее по возрастанию. Если в этом объединенном вариационном ряду объема $n_1 + n_2 = n$ окажется, что элементы одной выборки в целом больше элементов второй выборки, то понятно, что ни о какой однородности не может быть речи. Измеряют это преобладание при помощи числа инверсий: *если какой-либо элемент из выборки X встречается после какого-то элемента из выборки Y, то говорят, что имеет место одна инверсия.* Для упрощения изложения будем считать, что в общем вариационном ряду нет одинаковых элементов. Число инверсий будем обозначать через $\upsilon = U$, а число контринверсий через $\omega = V$.

Например, в последовательности $x\,y\,y\,x\,y\,x\,y\,y\,x\,x$ число инверсий $U = 15 = 0+2+3+5+5$. Максимальное число инверсий равно $U_{max} = 25$ и соответствует $y\,y\,y\,y\,y\,x\,x\,x\,x\,x$, минимальное $x\,x\,x\,x\,x\,y\,y\,y\,y\,y$ и равно $U_{min} = 0$. Очевидно, если число инверсий U близко к нулю, то элементы выборки X в совокупности меньше элементов выборки Y. Вопрос состоит в том, насколько можно объяснить случайностью ситуацию, когда $U > 0$.

Для проверки нулевой гипотезы $H_0 : \upsilon = \omega$, означающей, что число инверсий равно числу контринверсий, при односторонней альтернативной гипотезе $H_1 : \upsilon > \omega$ выбирается критическое число инверсий U_α такое, что если наблюдаемое число инверсий больше критического $U > U_\alpha$, то нулевая гипотеза отвергается

в пользу альтернативной с уровнем значимости α. Критическое значение U_α выбирается таким образом, чтобы в случае, когда нулевая гипотеза верна, количество перестановок с числом инверсий $U \geq U_\alpha$, не превышало $\theta_{\text{кр}} = \alpha\theta$, где θ — общее число возможных последовательностей, вычисляемое по формуле:

$$\theta = C_n^{n_1} = \frac{n!}{n_1!n_2!}.$$

Например, $n_1 = 5;\ n_2 = 5,\ \theta = C_{10}^5 = 252$, $\theta_{\text{кр}} = 6.3$ при $\alpha = 0.025$. Подсчет начинают с последовательности с максимальным числом инверсий до тех пор, пока не наберется больше чем $\alpha\theta$ последовательностей. Количество инверсий в последней последовательности принимают за U_α.

1 .	y	y	y	y	y	x	x	x	x	x	$U_1 = 25$
2 .	y	y	y	y	x	y	x	x	x	x	$U_2 = 24$
3 .	y	y	y	x	y	y	x	x	x	x	$U_3 = 23$
4 .	y	y	y	y	x	x	y	x	x	x	$U_4 = 23$
5 .	y	y	x	y	y	y	x	x	x	x	$U_5 = 22$
6 .	y	y	y	x	y	x	y	x	x	x	$U_6 = 22$
7 .	y	y	y	y	x	x	x	y	x	x	$U_7 = 22$

В последней последовательности число инверсий равно 22, поэтому $U_{0.025} = 22$. Наблюдаемое число инверсий $U_* = 15 < 22 = U_{0.025}$, поэтому выборки можно считать однородными. При двустороннем критерии нулевая гипотеза отвергается, если

$$U > U_{\alpha/2} \ \text{ или } n_1n_2 - U > U_{\alpha/2}.$$

Вместо подсчета инверсий можно суммировать ранги (порядок элемента в вариационном ряду).

$$U = (r_1 - 1) + (r_2 - 2) + \ldots + (r_{n_1} - n_1) = \\ = \sum r_i - \frac{n_1(n_1+1)}{2}.$$

8.1.2 Критерий Манна-Уитни

Представим число инверсий в виде

$$U = \sum_{i=1}^{n_1} \sum_{k=1}^{n_2} Z_{ik}, \quad \text{где} \quad Z_{ik} = \begin{cases} 1 & \text{при } x_i > y_k, \\ 0 & \text{иначе.} \end{cases}$$

Если выборки однородны, то случайная величина

$$Z_{ik} = \begin{cases} 1, & \text{с вероятностью } 0.5, \\ 0, & \text{с вероятностью } 0.5. \end{cases}$$

В этом случае $\mathbf{E}Z_{ik} = \frac{1}{2}$, $\mathbf{E}U = \frac{n_1 n_2}{2}$. Для вычисления дисперсии $\mathbf{D}U$ рассмотрим случайную величину $U - \mathbf{E}U = \sum\limits_{i,k} \left(Z_{ik} - \frac{1}{2}\right)$.

$$\mathbf{D}U = \mathbf{E}(U - \mathbf{E}U)^2 = \sum_{i,k,j,l} \mathbf{E}\left(Z_{ik} - \frac{1}{2}\right)\left(Z_{jl} - \frac{1}{2}\right).$$

Здесь нужно выделить четыре возможных случая:

- Слагаемые с индексами $i \neq j$ и $k \neq l$ равны нулю, так как элементы выборок независимы и

$$\mathbf{E}\left(Z_{ik} - \frac{1}{2}\right)\left(Z_{jl} - \frac{1}{2}\right) = \mathbf{E}\left(Z_{ik} - \frac{1}{2}\right)\mathbf{E}\left(Z_{jl} - \frac{1}{2}\right) = 0.$$

- Пусть $i = j$, $k = l$. Тогда

$$\mathbf{E}\left(Z_{ik} - \frac{1}{2}\right)^2 = \frac{1}{2}\left(1 - \frac{1}{2}\right)^2 + \frac{1}{2}\left(0 - \frac{1}{2}\right)^2 = \frac{1}{4}.$$

- Пусть $i = j$, $k \neq l$. В этом случае рассматриваются три индекса i, k, l, из которых индекс i относится к переменной из выборки X, а индексы k и l – к переменной из выборки Y. В случае однородных выборок возможны шесть равновероятных вариантов расположения этих трех элементов.

	индексы	порядок	Z_{ik}	Z_{il}	$(Z_{ik}-0.5)(Z_{il}-0.5)$
1.	(ikl)	$X_iY_kY_l$	0	0	$1/4$
2.	(ilk)	$X_iY_lY_k$	0	0	$1/4$
3.	(kil)	$Y_kX_iY_l$	1	0	$-1/4$
4.	(kli)	$Y_kY_lX_i$	1	1	$1/4$
5.	(lik)	$Y_lX_iY_k$	0	1	$-1/4$
6.	(lki)	$Y_lY_kX_i$	1	1	$1/4$

$$\mathbf{E}\left(Z_{ik}-\frac{1}{2}\right)\left(Z_{il}-\frac{1}{2}\right)=$$
$$=\frac{1}{6}\left(\frac{1}{4}+\frac{1}{4}-\frac{1}{4}+\frac{1}{4}-\frac{1}{4}+\frac{1}{4}\right)=\frac{1}{12}.$$

- В случае $i\neq j,\ k=l$ поступаем аналогично и получаем

$$\mathbf{E}\left(Z_{ik}-\frac{1}{2}\right)\left(Z_{jk}-\frac{1}{2}\right)=\frac{1}{12}.$$

Следовательно,

$$\mathbf{D}U=\left(\frac{1}{4}\right)n_1n_2+\left(\frac{1}{12}\right)n_1(n_2^2-n_2)+\left(\frac{1}{12}\right)n_2(n_1^2-n_1)=$$
$$=\frac{n_1n_2}{12}(n_1+n_2+1).$$

Итак получаем, что число инверисий $U=R_1-\frac{n_1(n_1+1)}{2}$ имеет математическое ожидание, равное $\mathbf{E}U=\frac{n_1n_2}{2}$ и дисперсию, равную $\mathbf{D}U=n_1n_2(n_1+n_2+1)/12$. При небольших объемах выборок для проверки гипотезы однородности используют по готовым таблицам критические значения числа инверсий. Нормальное приближение оказывается достаточно точным при $n_1>3, n_1+n_2>20$, поэтому для проверки нулевой гипотезы используется статистика

$$Z=\frac{R_1-\frac{n_1(n_1+1)}{2}-\frac{n_1n_2}{2}}{\sqrt{\frac{n_1n_2}{12}(n_1+n_2+1)}}\quad\sim\quad\mathcal{N}(0;1)\,. \tag{75}$$

8.1.3 Непараметрический критерий равенства дисперсий

Критерий для проверки нулевой гипотезы $H_0 : \sigma_1^2 = \sigma_2^2$ применяется следующим образом. Объединенная выборка объема $n_1 + n_2$ упорядочивается в порядке возрастания и отмечается принадлежность каждого элемента к той или иной выборке. Ранги присваиваются по правилу: минимальное значение получает ранг 1, два наибольших получают ранги 2 и 3, ранги 4 и 5 получают следующие наименьшие значения и так далее. Схема расстановки рангов выглядит следующим образом:

$$1, 4, 5, 8, 9, \ldots, 7, 6, 3, 2 \tag{76}$$

Смысл этого правила заключается в том, что выборка, дисперсия в которой меньше, будет сконцентрирована в середине объединенной выборки, следовательно, ее сумма рангов будет существенно выше.

Каждому из совпадающих по величине элементов присваивается ранг, равный среднему арифметическому рангов одинаковых элементов. При $n_1 > 10, n_2 > 10$ статистика критерия определяется по формуле:

$$Z = \frac{\left| R_2 - \frac{n_2(n_1+n_2+1)}{2} \right| - \frac{1}{2}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}}, \tag{77}$$

где R_2 – сумма рангов для выборки меньшего объема ($n_2 \leq n_1$), статистика $Z \sim \mathcal{N}(0; 1)$ нормальна при справедливости H_0.

ПРИМЕР. В табл.3 представлены данные о естественном приросте населения в 1997 году (ЕП) в разных странах. В 16 европейских странах медиана ЕП равна 0.7 ($\bar{x}_1 = -0.106$, $\hat{\sigma}_1 = 3.06$), в остальных пяти странах медиана ЕП равна 5.7 ($\bar{x}_1 = 8.56$, $\hat{\sigma}_2 = 8.47$). Согласно критерию Манна-Уитни, это различие значимо ($p = 0.002$) и не может быть объяснено случайностью. Сумма рангов $R_2 = 33$, $n_2 = 5$, $Z = 1.77$, $p = 0.04$, следовательно, гипотеза равенства дисперсий отвергается в пользу альтернативной, означающей меньший разброс наблюдений в Европе.

8.1.4 Критерий Краскела-Уоллиса

Критерий Краскела-Уоллиса является непараметрическим аналогом однофакторного дисперсионного анализа и применяется для проверки однородности нескольких независимых выборок. Пусть имеются наблюдения некоторого признака в r группах, $n_1 + \ldots + n_r = n$. Метод заключается в определении рангов каждого из элементов смешанной выборки и вычислении суммы рангов $S_i,\ i = 1, \ldots, r$, по каждой выборке. Однородность по группам проверяется при помощи статистики

$$\chi^2 = \frac{12}{n(n+1)} \sum_{i=1}^{r} \frac{S_i^2}{n_i} - 3(n+1)\,,$$

которая при H_0 имеет распределение $\chi^{\mathbf{2}}(r-1)$.

Если в объединенной выборке имеются m групп одинаковых наблюдений размера $t_i,\ i = 1, 2, \ldots, m$, то ранги, относящиеся к ним, усредняются, вычисляются поправочные коэффициенты $T_i = t_i^3 - t_i$, а значение статистики H меняется на

$$H' = \frac{H}{T}, \quad \text{где} \quad T = 1 - \sum_{i=1}^{m} \frac{T_i}{n^3 - n}\,.$$

8.1.5 Медианный критерий

Медианный тест применяется для проверки однородности нескольких независимых выборок. Пусть меются наблюдения в r группах, $n_1 + \ldots + n_r = n$. Через $x^{(1)} \le \ldots \le x^{(n)}$ обозначим объединенный вариационный ряд, n_{1i} и n_{2i} — число наблюдений в i-й группе соответственно больше и меньше медианы

$$M = \begin{cases} \frac{x^{(m)} + x^{(m+1)}}{2} & \text{при } n = 2m, \\ x^{(m+1)} & \text{при } n = 2m+1. \end{cases}$$

	Группы			
	1	$\ldots$	r	
$< M$	n_{11}	$\ldots$	n_{1r}	n_{1*}
$> M$	n_{21}	$\ldots$	n_{2r}	n_{2*}
	n_{*1}	$\ldots$	n_{*r}	n

При однородности выборок статистика

$$\chi^2 = \sum_{j=1}^{r} \sum_{i=1}^{2} \frac{(n_{ij} - \frac{n_{*j} n_{i*}}{n})^2}{\frac{n_{*j} n_{i*}}{n}}$$

имеет распределение хи-квадрат с $r-1$ степенями свободы. При значении $p = \{\chi^2 > \chi^2_*\}$, меньшем α, гипотеза однородности отвергается, и с уровнем значимости α внутригрупповые отклонения от медианы невозможно объяснить случайностью.

ПРИМЕР (продолжение). В четырех группах стран, отличающихся принадлежностью к Европе (E) и разным уровнем смертности от злокачественных новообразований (1 — высокая), имеется значимое различие по ЕП согласно критерию Краскела-Уоллиса ($p = 0.02$). В медианном критерии значимость равна $p = 0.06$.

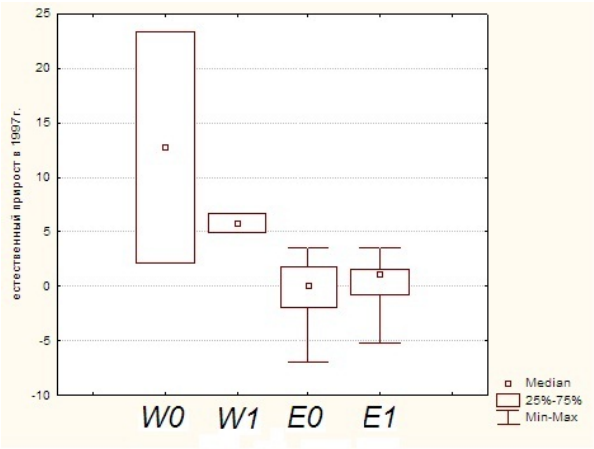

Рис. 6: Медианы ЕП в группах стран.

Попарные сравнения указывают, что наибольшим образом отличаются европейские страны от остальных при условии низкой смертности от злокачественных новообразований.

8.2 Критерии для зависимых выборок

8.2.1 Критерий знаков (л.11)

Примером зависимых выборок могут служить данные о наблюдениях за массой тела до и после диеты у 10 индивидов.

	1	2	3	4	5	6	7	8	9	10
Масса до диеты	68	80	92	81	70	79	78	66	57	76
Масса после диеты	60	84	87	79	74	71	72	67	57	70
Разность	8	−4	5	2	−4	8	6	−1	0	6

Из 9 индивидов, у которых произошли изменения в массе тела после диеты, шесть индивидов уменьшили массу тела, а трое даже увеличили. Один индивид остался с прежним весом.

Итак, имеется парная выборка (x_{1j}, x_{2j}), $j = 1, \ldots, n$. Рассмотрим разности $z_j = x_{2j} - x_{1j}$, $j = 1, \ldots, n$, в виде выборки наблюдений случайной величины ξ. Если выборки одинаково распределены, то вероятности положительных и отрицательных значений ξ должны быть одинаковы. Вследствие непрерывности случайных величин, $\xi = 0$ с нулевой вероятностью. Если одинаковые наблюдения встречаются из-за погрешности вычислений или ошибок округления, то их следует исключить из выборки. Для определенности будем считать успехом $\xi > 0$. Обозначим через $p = P\{\xi > 0\}$. Нулевая гипотеза имеет вид $H_0 : p = 0.5$.

Пусть из $n = n_+ + n_-$ элементов парной выборки имеется n_+ и n_- разностей со знаком $+$ и $-$, через ν_+ и ν_- обозначим случайное число плюсов и минусов соответственно.

При альтернативной гипотезе $H_1 : p > 0.5$ вычисляется вероятность получить n_- неудач или даже меньше:

$$\alpha_* = P\{\nu_- \le n_-\} = \sum_{k=0}^{n_-} P\{\nu_- = k\} = \frac{1}{2^n} \sum_{k=0}^{n_-} C_n^k.$$

При альтернативной гипотезе $H_1 : p < 0.5$ вычисляется ве-

роятность получить n_+ успехов или даже больше:

$$\alpha_* = P\{\nu_+ \geq n_+\} = \sum_{k=n_+}^{n} P\{\nu_+ = k\} = \frac{1}{2^n} \sum_{k=n_+}^{n} C_n^k.$$

При $H_1 : p \neq 0.5$ вычисляется значимость

$$\alpha_* = \frac{1}{2^{n-1}} \sum_{k=0}^{\min(n_-, n_+)} C_n^k.$$

Если $\alpha_* < \alpha$, то гипотеза о случайности в колебаниях знаков разностей отвергается с уровнем значимости α, иначе нет оснований отвергнуть эту гипотезу.

В примере, считая успехом повышение массы, получаем $\alpha_* = 0.25$. Поэтому H_0 не отвергается в пользу альтернативной $H_1 : p > 0.5$, число неудач мало, но это можно объяснить случайностью.

8.2.2 Критерий Вилкоксона для зависимых выборок

Также имеется парная выборка (x_{1j}, x_{2j}), $j = 1, \ldots, n$, и разности $z_j = x_{2j} - x_{1j}$. Выборку из модулей разностей $|z_j|$ проранжируем по возрастанию. Вычислим суммы рангов S_+ и S_-, соответствующие положительным и отрицательным разностям. Если все разности одного знака, то

$$\min\{S_+, S_-\} = 0, \quad \max\{S_+, S_-\} = \frac{n(n+1)}{2}.$$

При условии однородности зависимых выборок

$$E(S_-) = \frac{n(n+1)}{4},$$
$$D(S_-) = \frac{n(n+1)(2n+1)}{24}.$$

Таким образом, для проверки гипотезы однородности зависимых выборок используется статистика

$$Z = \frac{S_- - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}},$$

которая при справедливости нулевой гипотезы имеет стандартное нормальное распределение.

8.3 Критерий Фридмана

Этот критерий предназначен для проверки однородности нескольких зависимых выборок или для сопоставления условий эксперимента.

Имеется n индивидов, у которых измеряется значение одного и того же признака при c условиях. Предполагаем, что между полученными в разных условиях измерениями имеются случайные различия. Наблюдениям вида

$$\begin{array}{cccc} x_{11} & x_{12} & \dots & x_{1c} \\ x_{21} & x_{22} & \dots & x_{2c} \\ \vdots & \vdots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nc} \end{array}$$

сопоставляются ранги по строкам

$$\begin{array}{cccc} r_{11} & r_{12} & \dots & r_{1c} \\ r_{21} & r_{22} & \dots & r_{2c} \\ \vdots & \vdots & \dots & \vdots \\ r_{n1} & x_{r2} & \dots & r_{nc} \end{array}$$

и вычисляеются суммы рангов $R_i = \sum\limits_{j=1}^{n} r_{ij}$, средний ранг $\bar{R}_i = \frac{1}{n} R_i$, соответствующий условию $i = 1, 2, \dots, c$, и общий средний ранг вида $R = \frac{c+1}{2}$. Статистика критерия имеет вид

$$S = \frac{12n}{c(c+1)} \sum_{i=1}^{c} (\bar{R}_i - R)^2,$$

которая при $n \geq 13$, $c \geq 20$ имеет приближенное распределение хи-квадрат с $c-1$ степенями свободы. При меньших значениях параметров используются специальные вычисления.

Таблица 3: Пример данных для проверки гипотез однородности независимых выборок или независимости категориальных признаков.

Страна	Ест.пр.97	Европа	выс.см-ть
Россия	-5,2	1	1
Австралия	6,7	0	1
Австрия	0,5	1	0
Бельгия	1,2	1	0
Болгария	-7	1	1
Великобритания	1,6	1	1
Венгрия	-3,8	1	1
Германия	-0,8	1	1
Дания	1,5	1	1
Италия	-0,4	1	1
Канада	4,9	0	1
Мексика	23,4	0	0
Нидерланды	3,5	1	1
Норвегия	3,5	1	0
Польша	0,9	1	1
Румыния	-1,9	0	0
США	5,7	0	1
Финляндия	1,8	1	0
Франция	3,3	1	1
Швеция	-0,4	1	0
Япония	2,1	0	0

9 Проверка статистических гипотез

9.1 Простые гипотезы

9.1.1 Вероятность выигрыша в игре в кости

Игральными костями служили кости животных – астрагалы, которые при бросании могли падать на четыре стороны. Эти стороны как-то нумеровались, но единой системы не было. В одной из игр в древней Греции бросали одновременно четыре астрагала. Выигрышным броском считался тот, при котором выпадали разные стороны; такой бросок назывался „Венерой“. В археологических раскопках, начиная с V тысячелетия до н.э. (возможно, ранее) среди найденных костей астрагалы встречаются в несколько десятков раз чаще, чем другие кости.

После многократных исследований различных астрагалов были получены следующие частоты выпадения различных сторон. Частота выпадения широкой стороны A с углублением примерно равна 0.39, следующей по величине B равна 0.37; частоты выпадения двух оставшихся сторон C и D равны 0.12.

При подбрасывании четырех костей одновременно общее число исходов равно $4^4 = 256$. Например,

$$P(AAAA) = 0.39^4 = 0.0231,$$
$$P(ABCD) = 0.39 \cdot 0.37 \cdot 0.12 \cdot 0.12 = 0.002.$$

Всего броску "Венера" соответствуют $24 = 4{\cdot}3{\cdot}2{\cdot}1$ варианта перестановок названных сторон на четырех астрагалах:

$ABCD$	$ABDC$	$ACBD$	$ACDB$	$ADBC$	$ADCB$
$BACD$	$BADC$	$BCAD$	$BCDA$	$BDAC$	$BDCA$
$CABD$	$CADB$	$CBAD$	$CBDA$	$CDAB$	$CDBA$
$DABC$	$DACB$	$DBAC$	$DBCA$	$DCAB$	$DCBA$

Суммированием соответствующих вероятностей получаем вероятность броска "Венера" , равную $24 \cdot 0.002 = 0.048 \approx 0.05$.

9.1.2 Принцип маловероятных событий

Для проверки статистической гипотезы H_0 относительно параметров θ или других свойств генеральной совокупности с функцией распределения $F(x|\theta)$ случайной величины ξ используются выборочные наблюдения $x_1, \ldots, x_n$. Конкурирующую или альтернативную гипотезу будем обозначать через H_1.

Гипотеза называется простой, если она однозначно определяет функцию распределения, например, $H_0 : \theta = \theta_0$, иначе сложной, например, $H_0 : \theta > \theta_0$.

Правило, согласно которому отвергается гипотеза, называется *статистическим критерием,* а используемая для проверки гипотезы функция от выборочных наблюдений $x_1, \ldots, x_n$ называется *статистикой критерия.*

Проверка гипотез основана на *принципе маловероятных событий*, согласно которому события, вероятность которого мала, считаются невозможными. Сложилось так, что в качестве „малой“ вероятности принимают вероятность $\alpha = 0.05$ выигрыша в игре в кости. Гипотезу отвергают тогда, когда наблюдаемое значение статистики попадает в *критическую область* V, которую выбирает так, чтобы вероятность $P(V|H_0)$ была мала,

$$P(V|H_0) \leq \alpha\,.$$

ПРИМЕР. На $n = 25$ автомобилях с усовершенствованным двигателем средний расход бензина составил $\bar{x} = 9.3$ л на 100 км. Считая выборку нормальной $\mathcal{N}(\mu, \sigma)$ с дисперсией $\sigma^2 = 4$, выясним, нельзя ли наблюдаемое улучшение технической характеристики считать случайным. Проверим гипотезу $H_0 : \mu = 10$ о том, что расход топлива не изменился, В качестве альтернативной рассмотрим гипотезу $H_1 : \mu < 10$.

Используя свойства выборочного среднего $\bar{x}$, рассмотрим статистику критерия $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ с критической областью $V = (-\infty; -1.645]$, о которой известно, что $P(V) = 0.05$. Подставляя значения параметров и $\bar{x} = 9.3$, получаем значение статистики $Z = -1.75 \in V$, следовательно, гипотеза H_0 отвергается в пользу альтернативной H_1, уменьшение топлива значимо и не может быть объяснено случайностью.

9.1.3 Ошибки первого и второго рода, мощность

Ошибка первого рода возникает в случае, когда отвергается верная гипотеза, $\alpha = P(V|H_0)$. *Ошибка второго рода* возникает тогда, когда не отвергается ложная гипотеза $\beta = P(\bar{V}|H_1)$. *Мощностью* называется вероятность $P(V|H_1) = 1 - \beta$.

Критерий, имеющий среди критериев, удовлетворяющих условию $P(V|H_0) < \alpha$, наибольшую мощность $P(V|H_1)$, называется *наиболее мощным* относительно H_1. Критерий для проверки H_0 называется *несмещенным,* если $\forall H_1 \; P(V|H_0) \leq P(V|H_1)$.

Например, для проверки гипотезы $H_0 : \mu = 10$, при $\sigma = 2$, $n = 25$ с критической областью V, соотвествующей неравенству

$$\bar{x} < \mu - Z_{0.95}\frac{\sigma}{\sqrt{n}} = 9.342,$$

где $Z_{0.95}$ — квантиль $\mathcal{N}(0,1)$, ошибка первого рода равна

$$P\{\bar{x} < 9.342|\mu = 10\} =$$
$$= P\left\{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < \frac{9.342-\mu}{\sigma/\sqrt{n}}|\mu = 10\right\} = \Phi(-1.645) = 0.05\,.$$

Сложная гипотеза $H_1 : \mu < 10$ рассматривается как множество простых. Если $H_1 : \mu = 9$, тогда $1 - \beta = P\{\bar{x} < 9.342|\mu = 9\} =$

$$= P\left\{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < \frac{9.342-\mu}{\sigma/\sqrt{n}}|\mu = 9\right\} = \Phi(0.855) = 0.804\,.$$

Рассмотрим еще два критерия: $9.975 < \bar{x} < 10.025$ и $\bar{x} > 10.658$ с ошибками первого рода, также равными 0.05, и вычислим мощности при альтернативной гипотезе $H_1 : \mu = 9$.

$$\alpha_1 = \Phi\left(\frac{10.025-10}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{9.975-10}{\sigma/\sqrt{n}}\right) = 0.05,$$
$$1-\beta_1 = \Phi\left(\frac{10.025-9}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{9.975-9}{\sigma/\sqrt{n}}\right) = 0.0022,$$
$$\alpha_2 = 1 - \Phi\left(\frac{10.658-10}{\sigma/\sqrt{n}}\right) = 0.05,$$
$$1-\beta_2 = 1 - \Phi\left(\frac{10.658-9}{\sigma/\sqrt{n}}\right) = 1.7\cdot 10^{-5}\,.$$

Итак, критерий $\bar{x} < 9.342$ оказывается наиболее мощным, но смещенным, так как $P(V|H_1) < 0.05$ при $H_1 : \mu > 10$. Несмещенным является критерий $|\bar{x} - \mu| > Z_{0.975}\frac{\sigma}{\sqrt{n}}$ с критической областью $V = (-\infty; 9.216] \cup [10.784; +\infty)$. Например, при $H_1 : \mu = 9$ и $H_1 : \mu = 11$ имеем $P(V) = 0.705$.

9.1.4 Критерий отношения правдоподобия

Задача заключается в поиске наиболее мощного критерия, то есть в поиске критической области V, для которой $P(V|H_1)$ будет наибольшей при условии

$$P(V|H_0) < \alpha. \tag{78}$$

Обозначим через E пространство непрерывных переменных $x_1, \ldots, x_n$, и функции множеств $P(B) = P(B|H_0)$ и $P'(B) = P(B|H_1)$ определяются непрерывными плотностями

$$f(X) = f(x_1, \ldots, x_n) \text{ и } g(X) = f(x_1, \ldots, x_n)\,.$$

Если $f(X) = 0$, где $X \in T \subset E$, то при присоединении T к V условие (78) не изменится, а мощность $P'V$ может только увеличиться. Следовательно, на множестве $E \setminus T$ плотность f не равна нулю, и функция

$$U = U(X) = \frac{g}{f} \tag{79}$$

является непрерывной функцией,

$$P\{U < v\} = P\{g < vf\}\ .$$

Предложение 13 *Пусть* $\exists\ v > 0$, *такое что* $P\{g < vf\} = 1 - \alpha$. *Построим критическую область* V, *определяемую неравенством*

$$\frac{g}{f} \geq v\,.$$

Если другая область W *удовлетворяет условию (78), тогда* $P'W \leq P'V$.

Доказательство. Так как

$$\begin{gathered} P(V) = P(V \cap W) + P(V \setminus W) = \alpha, \\ P(W) = P(W \cap V) + P(W \setminus V) \leq \alpha, \\ \text{то } P(V \setminus W) \geq P(W \setminus V) \iff \\ \int\limits_{V \setminus W} f dX \geq \int\limits_{W \setminus V} f dX\,. \end{gathered} \tag{80}$$

В критической области $g \geq vf$. Следовательно,

$$P'(V) = P'(V \cap W) + P'(V \setminus W) = \\ = P'(V \cap W) + \int\limits_{V \setminus W} g dX \geq \\ \geq P'(V \cap W) + \int\limits_{V \setminus W} vf dX \,.$$

Отсюда, в силу (80), получаем

$$P'(V) \geq P'(V \cap W) + \int\limits_{W \setminus V} vf dX \,.$$

На множестве $W \setminus V$ для всех точек множества справедливо $g < vf$, поэтому

$$P'(V) \geq P'(V \cap W) + \int\limits_{W \setminus V} g dX = \\ = P'(V \cap W) + P'(V \cap W) = P'(W) \,.$$

Что и требовалось доказать.

Так как плотности $f(X)$ и $g(X)$ являются функциями правдоподобия гипотез H_0 и H_1, их отношение называется *отношением правдоподобия*. Только что полученный критерий отношения правдоподобия можно сформулировать так:

Гипотезу H_0 следует отвергнуть, если отношение правдоподобия

$$U = \frac{g(x)}{f(x)} \geq v.$$

Критическое значение v выбирается таким образом, чтобы вероятность ошибки первого рода равнялась $\alpha = P\{U \geq v\}$.

Этот критерий целесообразен в случае, когда имеется большая уверенность в том, что H_1 может быть верна.

9.1.5 Пример критерия отношения правдоподобия

Пусть $x_1, \ldots, x_n$ выборка над $\mathcal{N}(0,1)$, тогда

$$f(X) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}(x_1^2+\ldots+x_n^2)}.$$

Согласно альтернативной гипотезе H_1, предполагаем, что случайные величины $x_1, \ldots, x_n \sim \mathcal{N}(\mu, 1)$, где $\mu > 0$,

$$g(X) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}((x_1-\mu)^2+\ldots+(x_n-\mu)^2}.$$

Отношение правдоподобия равно

$$U = \frac{g}{f} = e^{\mu\sum_i x_i - \frac{n\mu^2}{2}}$$

и представляет собой монотонно возрастающую функцию от выборочного среднего $\bar{x} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$.

Гипотезу H_0 следует отвергнуть, когда $\bar{x} > c$. Критическое значение c выбирается таким образом, чтобы $P\{\bar{x} > c\} = \alpha$ при справедливости H_0. Следовательно, $c = Z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$. В случае $\mu > 0$ этот критерий не зависит от μ и является *равномерно наиболее мощным*.

Если в качестве конкурирующей гипотезы H_1 взять нормальное распределение с математическим ожиданием $\mu < 0$, тогда H_0 отвергается при $\bar{x} < -c$.

9.2 Сложные гипотезы

Рассмотрим гипотезу H_0 о том, что некоторая параметрическая точка θ принадлежит заданному подмножеству Θ из множества допустимых точек $\Theta \cup \bar{\Theta}$.

Например, при проверке гипотезы $H_0 : \mu = \mu_0$ в случае неизвестной дисперсии σ^2 имеется множество допустимых точек (μ, σ), где $\sigma > 0$, и Θ состоит из точек прямой $\mu = \mu_0$ в полуплоскости $\sigma > 0$.

Критерий для гипотезы H_0 представляет собой правило, по которому H_0 отвергается, если выборочное наблюдение x принадлежит некоторому критическому множеству V. Определим множество V так, чтобы мала была функция мощности

$$P(V,\theta) = P\{x \in V | \theta \in \Theta\},$$

и велика функция мощности

$$P'(V,\theta) = P\{x \in V | \theta \in \bar{\Theta}\}.$$

В некоторых случаях удается найти критическое множество V такое, чтобы $P(V,\theta) = \alpha$ для всех $\theta \in \Theta$. В таком случае говорят, что V *подобно пространству выборок* относительно множества Θ. Соответствующий критерий имеет уровень α.

Если удается найти такой критерий W уровня α, что для любого другого критерия V уровня α выполняется неравенство

$$P'(W,\theta) \geq P'(V,\theta),$$

где $\theta \in \bar{\Theta}$, то такой критерий называется *равномерно наиболее мощным* для гипотезы H_0 относительно множества допустимых гипотез $\bar{\Theta}$.

Если для критерия V уровня α имеем

$$P(V,\theta) > \alpha, \quad P'(V,\theta) > \alpha$$

для всех допустимых $\theta \in \Theta \cup \bar{\Theta}$, то такой критерий называется *несмещенным.*

Наиболее мощным несмещенным критерием является критерий W уровня α, что

$$P'(W,\theta) \geq P'(V,\theta),$$

где $\theta \in \bar{\Theta}$, для всех несмещенных критериев V.

10 Минимальный объем выборки

10.1 Задача о доле брака

ПРИМЕР. В партии из $n = 100$ изделий оказалась доля брака равна $h = 0.1$. Нужно построить доверительный интервал для вероятности брака p и определить, какой минимальный объем выборки следует взять, чтобы с вероятностью $1-\alpha$ можно было утверждать, что доля бракованных деталей во всей партии отличается не более чем на $\epsilon = 0.01$.

Пусть через $x_1, \ldots, x_n$ обозначена выборка наблюдений над случайной величиной

$$\xi = \begin{cases} 1 & \text{с вероятностью } p, \\ 0, & q, \end{cases}$$

где $q = 1 - p$, представляющая собой последовательность из нулей и единиц, означающих соответственно неудачу и успех, $H_0 : p = p_0$, $H_1 : p = p_0 + \epsilon$.

Эффективной оценкой вероятности успеха в одном испытании является относителная частота

$$\hat{p} = h = \bar{x} \sim \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right),$$

которая асимптотически имеет нормальное распределение по теореме Муавра-Лапласа. Из асимптотической $\mathcal{N}(0,1)$ нормальности статистики

$$Z = \frac{h - p}{\sqrt{pq}}\sqrt{n}$$

имеем, что при больших n приблизительно с вероятностью $1 - \alpha$ выполняется неравенство

$$h - z_{1-\frac{\alpha}{2}}\sqrt{\frac{pq}{n}} < p < h + z_{1-\frac{\alpha}{2}}\sqrt{\frac{pq}{n}},$$

где через $z_{1-\frac{\alpha}{2}}$ обозначена квантиль стандартного нормального распределения. Заменяя значения p и q в границах неравенства

соответствующими оценками h и $1-h$, получаем приближенный доверительный интервал для вероятности успеха p в схеме Бернулли

$$h - z_{1-\frac{\alpha}{2}}\sqrt{\frac{h(1-h)}{n}} < p < h + z_{1-\frac{\alpha}{2}}\sqrt{\frac{h(1-h)}{n}}\,.$$

В нашей задаче оказывается $0.041 < p < 0.159$. Выясним, при каких n должно выполняться неравенство

$$|h-p| < z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} \le \epsilon\,.$$

Нетрудно убедиться, что

$$n \ge z^2_{1-\frac{\alpha}{2}}\frac{p(1-p)}{\epsilon^2}|_{p=h} \approx 3457\,.$$

10.2 Критерий равенства

Второй тип задач, связанный с определением необходимого объема выборки, относится к планированию эксперимента и чаще всего встречается при организации клинических исследований. Пусть необходимо найти необходимый объем выборки для сравнения двух частот с предполагаемым расстоянием 0.15, с уровнем значимости $\alpha = 0.05$ и с мощностью $1-\beta = 0.80$. Обозначим через $\epsilon = p_1 - p_2$. Дополнительное условие связано с пропорцией между объемами n_1 контрольной и n_2 референтной групп, $n_2 = \kappa n_1$.

$$H_0 : \epsilon = 0, \quad H_1 : |\epsilon| > 0.$$

Обозначим генеральную частоту в одном эксперименте через p_1, во втором через p_2, соответственно $q_i = 1 - p_i$, $i = 1, 2$ и построим таблицу сопряженности.

эксперимент	$-$	$+$	всего
1	n_{11}	n_{12}	$n_{1\cdot} = n_1$
2	n_{21}	n_{22}	$n_{2\cdot} = n_2$
оба	$n_{\cdot 1}$	$n_{\cdot 2}$	n

По сути у нас имеются две независимые выборки $x_{11}, \ldots, x_{1n_1}$ и $x_{21}, \ldots, x_{2n_2}$ из нулей и единиц, частоты выражаются через выборочные средние: $\hat{p}_1 = \frac{n_{12}}{n_{1\cdot}} = \bar{x}_1$, $\hat{p}_2 = \frac{n_{22}}{n_{2\cdot}} = \bar{x}_2$. Критерий равенства частот основан на статистике

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - \epsilon}{\sigma\sqrt{\frac{1}{n_{1\cdot}} + \frac{1}{n_{2\cdot}}}} \sim \mathcal{N}(0,1)\,.$$

Обозначим для краткости через $\zeta = \bar{x}_1 - \bar{x}_2$, $S = \sigma\sqrt{\frac{1}{n_{1\cdot}} + \frac{1}{n_{2\cdot}}}$ и построим критическую область. В предположении $\epsilon = 0$

$$P\{|Z| > z_{1-\frac{\alpha}{2}}\} = \alpha \quad \Longleftrightarrow \quad P\{|\zeta| > z_{1-\frac{\alpha}{2}} S\} = \alpha\,.$$

При H_1 имеем $\zeta \sim \mathcal{N}(\epsilon, S)$, и вероятность попадания в доверительную область равна ошибке второго рода β,

$$\Phi\left(\frac{z_{1-\frac{\alpha}{2}} S - \epsilon}{S}\right) - \Phi\left(\frac{-z_{1-\frac{\alpha}{2}} S - \epsilon}{S}\right) = \beta \quad \Longleftrightarrow$$
$$\Phi\left(z_{1-\frac{\alpha}{2}} - \frac{\epsilon}{S}\right) - \Phi\left(-z_{1-\frac{\alpha}{2}} S - \frac{\epsilon}{S}\right) = \beta\,.$$

При известном значении σ последнее уравнение лучше всего решить численным образом относительно n_1. Однако, если пренебречь одним из слагаемых, то n_1 можно выразить формально.

$$\beta \approx \Phi\left(z_{1-\frac{\alpha}{2}} - \frac{|\epsilon|}{S}\right) \quad \Longleftrightarrow \quad z_{1-\frac{\alpha}{2}} - \frac{|\epsilon|}{S} = z_\beta\,.$$

Таким образом, $S^2(z_{1-\frac{\alpha}{2}} - z_\beta)^2 = \epsilon^2$, и при $n_2 = \kappa n_1$ получаем

$$n_1 = \frac{\sigma^2(1+\kappa^{-1})(z_{1-\frac{\alpha}{2}} - z_\beta)^2}{\epsilon^2}\,.$$

Остается проблема — чему равно σ^2. Максимально возможная дисперсия бернуллиевской случайной величины достигается в случае вероятности успеха 0.5, поэтому при $\sigma^2 = 0.25$ и $\kappa = 1$ имеем

$$n_1 = \frac{(z_{1-\frac{\alpha}{2}} - z_\beta)^2}{2\epsilon^2}\,.$$

Любые сведения о вероятности p_1 позволяют уменьшить необходимый объем выборки. При $\alpha = 0.05$, $\beta = 0.2$, $\epsilon = 0.15$ объем одинаковых выборок равен $n_1 = n_2 = 174$.

10.3 Критерий возрастания (неубывания)

Обозначим через $\varepsilon = \mu - \mu_0$. Введем нулевую и альтернативную гипотезы:

$$H_0 : \varepsilon \leq \delta, \quad H_1 : \varepsilon > \delta,$$

где $\delta > 0$ уровень возрастания. Предположим, что дисперсия σ^2 известна. Рассмотрим случайную величину $\bar{x}$ со стандартным отклонением $S = \frac{\sigma}{\sqrt{n}}$. При нулевой гипотезе $\mathbf{E}\bar{x} = \mu < \mu_0 + \delta$, а при альтернативной $\mathbf{E}\bar{x} = \mu > \mu_0 + \delta$. Найдем критическую область V_k. Мы отвергаем нулевую гипотезу с уровнем значимости α при

$$\frac{\bar{x} - \mu_0 - \delta}{\sigma/\sqrt{n}} > z_{1-\alpha} \quad \text{или} \quad \bar{x} > \mu_0 + \delta + z_{1-\alpha} S,$$

где z_α квантиль стандартного нормального распределения, то есть такое значение, что для стандартно нормально распределенной случайной величины $P\{\xi < z_\alpha\} = \alpha$.

Если верна альтернативная гипотеза $\varepsilon = \mu - \mu_0 > \delta$, то

$$P(V_k|H_1) = 1 - \Phi\left(\frac{(\mu_0 + \delta + z_{1-\alpha} S) - \mu}{S}\right) = \\ = \Phi\left(-z_{1-\alpha} + \frac{\varepsilon - \delta}{S}\right) = 1 - \beta\,,$$

где β ошибка второго рода. Отсюда

$$\frac{\varepsilon - \delta}{S} - z_{1-\alpha} = z_{1-\beta}, \iff \sqrt{n}(\varepsilon - \delta) = \sigma(z_{1-\alpha} + z_{1-\beta}), \\ n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{(\varepsilon - \delta)^2}\,.$$

Если нужен критерий непревосходства, то нулевая гипотеза имеет вид $H_0 : \varepsilon > \delta$, а критическая область задается соотношением

$$\frac{\bar{x} - \mu_0 - \delta}{S} < z_\alpha\,, \quad \bar{x} < z_\alpha S + \delta + \mu_0\,.$$

Если верна альтернативная гипотеза $\varepsilon = \mu - \mu_0 < \delta$, то

$$P(V_k|H_1) = \Phi\left(\frac{(z_\alpha S + \delta + \mu_0) - \mu}{S}\right) = \\ = \Phi\left(-z_{1-\alpha} + \frac{\delta - \varepsilon}{S}\right) = 1 - \beta\,,$$

где β ошибка второго рода. Отсюда

$$\frac{\delta - \varepsilon}{S} - z_{1-\alpha} = z_{1-\beta},\ \Longleftrightarrow\ \sqrt{n}(\delta - \varepsilon) = \sigma(z_{1-\alpha} + z_{1-\beta}),$$
$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2\sigma^2}{(\delta - \varepsilon)^2}\,.$$

10.4 Критерий эквивалентности

Возможно уменьшение объема выборки за счет введения так называемого предела эквивалентности δ, используемого в выражении для нулевой и альтернативной гипотез:

$$H_0 : |\epsilon| \geq \delta, \quad H_1 : |\epsilon| < \delta\,.$$

Также необходимо найти объем выборки для сравнения двух частот в предположении истинного различия между частотами меньше $\epsilon = 0.05$ с пределом эквивалентности $\delta = 0.01$, $\alpha = 0.05$, $\beta = 0.20$. Считаем $S = \frac{\sigma}{\sqrt{n}}$. Из $H_0 : \varepsilon > \delta$ и $H_0 : -\varepsilon > \delta$ (иначе $H_0 : \varepsilon < -\delta$) получим критическую область, задаваемую неравенствами соответственно

$$\frac{\bar{x} - \mu_0 - \delta}{S} < -z_{1-\alpha},\ \ \frac{\bar{x} - \mu_0 - (-\delta)}{S} > z_{1-\alpha}\,.$$

$$\mu_0 - \delta + z_{1-\alpha}S < \bar{x} < \mu_0 + \delta - z_{1-\alpha}S\,.$$

Выражение для мощности как вероятности попасть в критическую область, когда верна альтернативная, имеет вид:

$$\Phi\left(\frac{\mu_0+\delta-z_{1-\alpha}S-\mu}{S}\right)-\Phi\left(\frac{\mu_0-\delta+z_{1-\alpha}S-\mu}{S}\right)=$$
$$=\Phi\left(-z_{1-\alpha}+\frac{\delta-\epsilon}{S}\right)+\Phi\left(-z_{1-\alpha}+\frac{\epsilon+\delta}{S}\right)-1>$$
$$>2\Phi\left(-z_{1-\alpha}+\frac{\delta-|\epsilon|}{S}\right)-1>1-\beta\,.$$

Отсюда заключаем, что необходимый объем выборки можем получить из уравнения

$$-z_{1-\alpha}+\frac{\delta-|\epsilon|}{\sigma/\sqrt{n}}=z_{1-\beta/2},\ n=\frac{\sigma^2(z_{1-\alpha}+z_{1-\beta/2})^2}{(\delta-|\varepsilon|)^2}\,.$$

Задание 1

1. Промоделировать величину выигрыша в 50-ти кратном повторении эксперимента: из колоды в 36 карт наугад выбирается карта, и если это туз или какая-то другая карта (черная масть, пики, шестерка или семерка, не дама и т.д.), например, вероятность $= 1/9$, то выигрыш a, если нет, выигрыш $-b$, то есть проигрыш b. Лучше подобрать такие a и b, чтобы средний выигрыш был равен нулю. Можно это сделать с помощью Excel при помощи функции „=ЕСЛИ(СЛЧИС() $< ; a; -b$)“.

2. Проделать опыт $n = 100$ раз и вычислить характеристики распределения: математическое ожидание, дисперсию, стандартное отклонение, вариацию, рассеяние, медиану, моду, асимметрию, эксцесс; построить гистограмму, эмпирическую функцию распределения. Можно ли считать распределение случайного суммарного выигрыша нормальным? При таком предположении вычислить вероятность выигрыша, равного $10 \cdot b$, используя функцию нормального распределения. Сравнить теоретический результат с наблюдаемой частотой такого события.

3. Вычислить асимметрию и эксцесс а) равномерного распределения, б) гамма распределения с параметром формы больше единицы.

Задание 2.1

Промоделировать N раз нормально распределенную выборку объема n при помощи функции $rnorm()$, для каждой выборки оценить среднее значение при помощи функции $mean()$ и стандартное отклонение $sd()$. Проверить несмещенность, состоятельность и эффективность оценок среднего и дисперсии.

Задание 2.2

1) Построить оценки параметров экспоненциального распределения по ММ и ММП. Применить оценик к модельной выборке. Для моделирования можно использовать функцию $rgamma()$

при параметре $shape = 1$.

2) Иногда для оценки параметра λ в распределении Пуассона используется оценка в виде среднего арифметического среднего и несмещенной выборочной дисперсии. Будет ли эта оценка лучше оценки в виде выборочного среднего?

Задание 3

1) Проверить гипотезу согласия с нормальным законом распределения суммарного выигрыша из первого задания, используя гипотетические параметры.

2) Далее оценить по выборке параметры распределения и с ними проверить гипотезу согласия.

3) Построить гистограмму и плотность нормального распределения с заданными о оцененными параметрами.

Задание 4 (примере данных по AAC)

1. Построить две таблицы сопряженности по двум параме качественных признаков. При необходимости некоторые градации можно объединить.

    - Депрессия и тревожность (depressed.mood.1, anxiety.1), тяга к алкоголю и слабость (craving.to.alcohol.1, weakness.1)
    - Депрессия и боль в сердце (depressed.mood.1, chest.pain.1), раздражительность и тремор (irritability.1 tremor.1 (01,2))
    - Депрессия и гиперимия (depressed.mood.1, hyperemia.1), слабость и бессонница (weakness.1, insomia.1)
    - Тревожность и галлюцинации (anxiety.1,transient.hallusinations.1 (0,12)), потливость и гиперемия (sweating.1, hyperemia.1)
    - Подозрительность и диарея (suspiciousness.1, diarrhea.1), бессонница и анорексия (insomia.1, anoreksia.1)
    - Головная боль и боль в груди (headache.1 (0,12), chest.pain.1 (0,12)), депрессия и потливость (depressed.mood.1, sweating.1)

- Тошнота и жажда (vomiting.1 (0,12), thirst.1 (01,2)), анорексия и гиперемия (anoreksia.1, hyperemia.1)

2. Вычислить условные вероятности и указать на значимость их различия по критерию хи-квадрат и по точному критерию Фишера.

3. вычислить коэффициенты неопределенности. Упорядочить признаки по убыванию их неопределенности. В каких случаях можно говорить о значимой взаимосвязи между признаками, а в каких нет?

4. Применить критерии Мак-Немара и Кохрена для проверки значимых изменений во времени.

Задание 5.1
(данные http://mystatbook.narod.ru/minimumA.zip)

Применить критерии однородности для индексов тяжести в зависимости от занятости ($work$), пола, использования стимулирующих (st) и успокоительных (se) препаратов. Построить доверительные интервалы для средних.

Задание 5.2
(данные http://mystatbook.narod.ru/minimumA.zip)

Применить критерии однородности для индексов тяжести в зависимости от уровня образования ($educat$). Применить наведение контрастов.

Задание 6 (Данные AAC)

Исследовать корреляционную зависимость между двумя переменными из HR.1, SBP.1, DBP.1, MBP.1, SV.1, CO.1, SI.1, CI.1, TPR.1 и HR.9, SBP.9, DBP.9, MBP.9, SV.9, CO.9, SI.9, CI.9, TPR.9

1) в пределах одного дня (с одинаковым индексом)

2) одной и той же переменной в 1 и 9 день.

Оценить параметры линейной регрессии, коэффициент корреляции, проверить его значимость отклонения от нуля, построить двумерную диаграмму, линию регрессии и эллипс рассеяния.

Содержание