Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Учебное пособие

Алексеева Нина Петровна, Бурнаева Эльфия Гарифовна

Основные методы первичной обработки данных

Санкт-Петербург – 2023

Оглавление

Введение

Учебное пособие представляет собой практическое руководство по описательным задачам математической статистики. Представлены необходимые теоретические сведения, в большинстве случаев с доказательствами, выделенными для наглядности более мелким шрифтом. Рассматриваются свойства наиболее употребительных распределений, в том числе сложных, возможность их моделирования, вычисления выборочных характеристик, оценки параметров, а также проверки гипотез согласия с законом распределения. Показано, как можно выполнить первичную статистическую обработку в пакете R. Приведены примеры вычислений и варианты заданий для самостоятельной работы.

Данное пособие предназначено, прежде всего, для студентов или аспирантов, для которых разделы теории вероятностей и математической статистики не были профилирующими, но может быть полезно также научным сотрудникам или преподавателям – всем, кто заинтересован в получении эффективного инструмента для анализа и визуализации данных. Предполагается, что обучающимся ранее был прочитан курс теории вероятностей, но для более быстрой адаптации к теме в краткой форме приведены некоторые важные факты из этого курса. Часть разделов может быть использована в обучении методам статистической обработки информации для студентов нематематических специальностей.

1. Некоторые общие свойства случайных величин

Случайная величина - это величина, которая в результате опыта со случайным исходом принимает то или иное числовое значение, причем заранее неизвестно, какое именно. В случае конечного или счетного количества

исходов случайная величина ξ называется дискретной. Закон распределения такой величины задается в виде таблицы, в которой указано, какие значения может принимать случайная величина и с какими вероятностями, или аналитически в виде формулы, позволяющей вычислить эти вероятности. Вероятность события $\xi \leq x$ называется функцией распределения $F(x) = \mathbb{P}\{\xi \leq x\}$. Если существует плотность, то есть неотрицательная функция $f(x)$, такая что, $\int\limits_{-\infty}^{\infty} f(t)dt = 1$, и функцию распределения можно выразить как $F(x) = \int\limits_{-\infty}^{x} f(t)dt$, то такая случайная величина называется абсолютно непрерывной. В этом случае среднее значение или математическое ожидание вычисляется по формуле $\mathbb{E}\xi = \int\limits_{-\infty}^{\infty} xf(x)dx$, а в случае дискретного распределения вида $\xi : \begin{pmatrix} x_1 & x_2 & \dots & x_m & \dots \\ p_1 & p_2 & \dots & p_m & \dots \end{pmatrix}$ в виде $\mathbb{E}\xi = \sum\limits_{j=1}^{\infty} x_j p_j$. Напомним, что характеристика вариабельности случайной величины дисперсия вычисляется по формуле $\mathbb{D}\xi = \mathbb{E}(\xi - \mathbb{E}\xi)^2$. Начальные моменты k-го порядка случайной величины ξ условимся обозначать через $\alpha_k = \mathbb{E}\xi^k$, а центральные $\mu_k = \mathbb{E}(\xi - \alpha_1)^k$.

Иногда для доказательства некоторых теорем бывает полезно использовать производящую функцию, которая определяется для целочисленной случайной величины ξ, принимающей значения $0, 1, 2, \dots$ с вероятностями $p_0, p_1, p_2, \dots$, $\sum_{j=1}^{\infty} p_j = 1$, как

$$h(\nu) = \sum_{j=0}^{\infty} p_j \nu^j \,. \tag{1}$$

Нетрудно убедиться в том, что $h(1) = 1$,

$$\mathbb{E}\xi = h'(1)\,, \quad \mathbb{D}\xi = h''(1) + h'(1) - (h'(1))^2\,, \text{ так как} \tag{2}$$

$$h'(\nu) = \sum_{j=0}^{\infty} j p_j \nu^{j-1}|_{\nu=1} = \alpha_1 = \mathbb{E}\xi\,,$$

$$h''(\nu) = \sum_{j=0}^{\infty} j(j-1) p_j \nu^{j-2}|_{\nu=1} = \alpha_2 - \alpha_1,$$

$$\mathbb{D}\xi = \alpha_2 - \alpha_1^2 = (h''(1) + h'(1)) - (h'(1))^2\,.$$

Аналогично можно выразить моменты третьего и четвертого порядка и связанные с ними характеристики асимметрии $\gamma_1 = \frac{\mu_3}{\mu_2^{\frac{3}{2}}}$ и эксцесса $\gamma_2 = \frac{\mu_4}{\mu_2^2}$.

Вероятности распределения могут быть получены из производящей функции $h(\nu)$ при помощи дифференцирования $p_k = \frac{h^{(k)}(0)}{k!}$.

Одно из важных свойств производящих функций заключается в том, что производящая функция суммы двух независимых величин является произведением частных производящих функций. Это свойство справедливо и для производящей функции моментов, которая определяется для случайной величины ξ (необязательно целочисленной) как

$$g(t) = \mathbb{E}e^{t\xi} = \begin{cases} \int\limits_{-\infty}^{\infty} e^{tx} f(x)dx & \text{в непрерывном случае,} \\ \sum_{j=0}^{\infty} p_j e^{tj} & \text{в дискретном случае.} \end{cases} \tag{3}$$

Значение k-й производной производящей функции моментов в нуле совпадает с k-м начальным моментом. Действительно, $g'(t) = (\mathbb{E}e^{t\xi})' = \mathbb{E}\xi e^{t\xi}|_{t=0} = \mathbb{E}\xi$, $g''(t) = (\mathbb{E}\xi e^{t\xi})' = \mathbb{E}\xi^2 e^{t\xi}|_{t=0} = \mathbb{E}\xi^2$, и т.д. Иногда вместо этих функций используется характеристическая функция $\psi(t) = \mathbb{E}e^{it\xi}$.

Далее при рассмотрении некоторых видов распределения для вычисления дисперсии будет удобно воспользоваться свойством о том, что дисперсия суммы двух случайных величин ξ, η равна

$$\begin{aligned} D(\xi+\eta) = \mathbb{E}((\xi+\eta) - \mathbb{E}(\xi+\eta))^2 = \mathbb{E}((\xi - \mathbb{E}\xi) + (\eta - \mathbb{E}\eta))^2 = \\ = \mathbb{E}(\xi - \mathbb{E}\xi)^2 + \mathbb{E}(\eta - \mathbb{E}\eta)^2 + 2\underbrace{\mathbb{E}(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta)}_{cov(\xi,\eta)} \end{aligned}$$

Если величины ξ, η независимы, то их ковариация равна нулю,

$$cov(\xi,\eta) = \mathbb{E}(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta) = \mathbb{E}(\xi - \mathbb{E}\xi)\mathbb{E}(\eta - \mathbb{E}\eta) = 0,$$

и дисперсия суммы независимых величин равна сумме дисперсий.

2. Выборка, эмпирическое распределение, гистограмма

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 параметров, 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 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 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 x_{ji} *варианты* или значения признака ξ_i у индивида ω_j.

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 упорядоченную 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 $b \approx \frac{R}{k}$. После этого вычисляются частоты ν_i, равные количеству элементов выборки, попадающих в этот интервал. Очевидно, $\nu_1 + \ldots + \nu_k = n$. 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 со значениями $x_1, \ldots, x_n$ 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 любых* $x \in (-\infty; +\infty)$ *и* $\epsilon > 0$

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

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. Из теоремы Бернулли получаем

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

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

3. Характеристики выборочного распределения

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

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

Введем выражения для начальных и центральных выборочных моментов

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

Можно показать, что

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

Приведем вывод первых двух равенств, а третье по аналогии можно получить самостоятельно.

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

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 = \frac{1}{n}\sum_{i=1}^{n}(x_i^3 - 3x_i^2\bar{x} + 3x_i\bar{x}^2 - \bar{x}^3) = a_3 - 3\alpha_2\bar{x} + 3\bar{x}^3 - \bar{x}^3 = a_3 - 3\alpha_2\bar{x} + 2\bar{x}^3 .$$

Для того чтобы изучить свойства выборочных характеристик, нужно обратиться к некоторым теоретическим вопросам статистического оценивания параметров.

Определение 2. Пусть $x_1, \ldots, x_n$ — выборка из генеральной совокупности с функцией распределения $F(x, \theta)$, где θ параметр распределения. Точечной оценкой неизвестного параметра $\hat{\theta}$ называется приближенное значение этого параметра, полученного по выборке.

Очевидно оценка $\hat{\theta}$ есть значение некоторой функции элементов выборки, то есть $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$. Любую функцию от выборочных наблюдений называют статистикой. Для того чтобы иметь возможность сравнивать разные статистики, их рассматривают как функции некоторого случайного вектора, одной из реализаций которого является выборка. Так как закон распределения компонент этого случайного вектора $F(x, \theta)$ зависит от параметра θ, распределение статистики $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ также зависит от этого параметра. Для того чтобы выяснить, насколько одна оценка лучше другой, рассматривают разные свойства, связанные с их центром положения, рассеянием и поведением при увеличении объема выборки. На данном этапе нас интересует понятие несмещенности.

Определение 3. Оценка $\hat{\theta}_n$ называется *несмещенной,* если ее математическое ожидание совпадает с истинным значением параметра, т.е. $\mathbb{E}\hat{\theta}_n = \theta$.

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

Предложение 1. *Выборочный центральный момент* $m_2 = \frac{1}{n}\sum\limits_{j=1}^{n}(x_j - \bar{x})^2$ *второго порядка является смещенной оценкой дисперсии* σ^2*, то есть*

$$\mathbb{E}m_2 = \frac{\sigma^2(n-1)}{n} = \sigma^2 - \frac{\sigma^2}{n}. \tag{5}$$

Поэтому в качестве оценки дисперсии используют выражение

$$S^2 = \frac{n}{n-1}m_2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2\,. \tag{6}$$

Для того чтобы убедиться в этом, введем $y_i = x_i - \mu$ с математическим ожиданием $\mathbb{E}y_i = 0$ и дисперсией

$\mathbf{D}y_i = \sigma^2$, тогда $\bar{y} = \bar{x} - \mu$,

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 =$$

$$\text{из (4)} \quad = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2\,.$$

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

$$= \sigma^2 - \frac{1}{n^2}\sum_{i}\mathbb{E}y_i^2 = \sigma^2 - \frac{\sigma^2}{n}\,,$$

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

Стандартное отклонение выборочного среднего $\frac{S}{\sqrt{n}}$ иначе называют *ошибкой среднего.*

Выборочной *модой* унимодального распределения является элемент выборки mod, встречающийся с наибольшей частотой. Например, в выборке $1, 1, 1, 1, 2, 2, 2, 3, 3, 4$ модой является значение 1.

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 $med = x^{(k+1)}$. В данном примере $med = 2$.

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

$$\mathbb{P}\{\xi \leq x_P\} = P. \tag{7}$$

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.

Выборочный коэффициент асимметрии определяется как

$$g_1 = \frac{m_3}{m_2^{\frac{3}{2}}}\,.$$

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 2. *В случае симметричного закона распределения, когда для плотности распределения справедливо*

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

все нечетные моменты равны нулю.

Действительно,

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

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

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

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

При $g_1 > 0$ наблюдения слева сконцентрированы, а справа растянуты, и соответственно при $g_1 < 0$ наоборот.

Выборочный эксцесс $g_2 = \frac{m_4}{m_2^2} - 3$ используется для симметричных распределений в качестве метрики отклонения от нормального закона распределения, при котором эксцесс равен нулю. .

Предложение 3. *Пусть ξ распределена по нормальному закону с нулевым средним и единичной дисперсией, т.е. $\xi \sim \mathcal{N}(0, 1)$. Тогда $\mu_4 = 3$.*

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

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

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

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

При $h = 1$ из последнего уравнения получаем

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

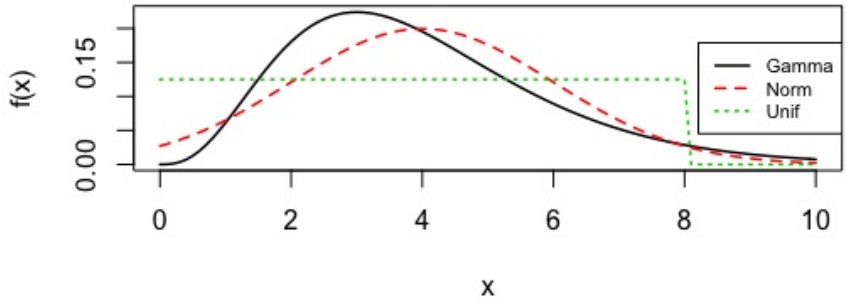

Рис. 1. Эксцеccs нормального, равномерного и гамма распределений равны соответственно 0, $-\frac{6}{5}$ и $\frac{6}{\lambda}$, где λ параметр формы.

Отсюда для $\xi \sim \mathcal{N}(0,1)$ эксцесс имеет вид $\gamma_2 = \frac{\mu_4}{\mu_2} - 3 = 0$. Нормально распределенная случайная величина ζ со средним μ и дисперсией σ^2 выражается через $\xi \sim \mathcal{N}(0,1)$ как $\zeta = \sigma\xi + \mu$, следовательно,

$$\mu_4(\zeta) = \sigma^4\mu_4(\xi),\ \mu_2(\zeta) = \sigma^2\mu_2(\xi),\ \gamma_2(\zeta) = \frac{\sigma^4\mu_4}{\sigma^4\mu_2} - 3 = 0\,.$$

4. Свойства некоторых распределений и их моделирование

Алгоритмы моделирования случайных величин с заданным законом распределения с помощью датчика псевдослучайных чисел — это отдельный раздел, изложение которых можно найти в учебниках по методу Монте-Карло. Далее под моделированием понимается генерирование случайных выборок в пакете R.

Практическое задание по первичной статистике предполагает изучение статистических свойств выборочных распределений, и очень полезно их изучать по выборкам, законы распределения которых известны. Лучше всего для этих целей подходят моделированные выборки с заданными значениями параметров. В связи с этим актуально повторить из курса теории вероятностей темы, связанные с основными законами распределения: биномиальным, Пуассона, равномерным, нормальным, гамма, Вейбулла и вспомнить способы вычисления их основных характеристик: математических ожиданий, дисперсии, коэффициентов асимметрии и эксцесса.

4.1. Биномиальное распределение

Итак, если известны значения $x_1, x_2, \ldots, x_N, \ldots$, которые принимает случайная величина ξ, а также вероятности $p_i = \mathbb{P}\{\xi = x_i\},\ i = 1, \ldots, N$, $p_1 + \ldots + p_N + \ldots = 1$, то говорят, что задан ее дискретный закон распределения

$$\xi : \begin{pmatrix} x_1 & x_2 & \ldots & x_N \\ p_1 & p_2 & \ldots & p_N \end{pmatrix}. \tag{8}$$

Например, при подбрасывании монеты возможны два исхода: успех или неудача, которые кодируются соответственно 1 и 0. Успех обычно ассоциируется с выпадением монеты строной "орел", которая выпадает с вероятностью $p = 0.5$. Неудача ассоциируется с выпадением стороны "решка", которая выпадает с вероятностью $q = 1 - p = 0.5$. Вероятность успеха p необязательно равна 0.5. Если при подбрасывании игрального кубика успехом считать выпадении шести очков, то $p = \frac{1}{6}$, $q = \frac{5}{6}$. Случайная величина, принимающая два значения с вероятностями p и $q = 1 - p$, имеет распределение Бернулли. Для краткости можно использовать обозначение $\xi \sim \beta(p, 1)$.

Обозначим через ξ случайное число выпадения шести очков игрального кубика в $n = 4$ независимых испытаниях с вероятностью выпадения шести очков (успех) $p = \frac{1}{6}$. Построим закон распределения этой случайной величины. Вычисление вероятностей $P\{\xi = k\},\ k = 0, 1, \ldots, 4$, отображено в таблице 1, где через $q = 1 - \frac{1}{6} = \frac{5}{6}$ обозначена вероятность невыпадения шести очков (неудача), через C_n^k число сочетаний по k из n элементов, которое вычисляется по формуле

$$C_n^k = \frac{n!}{k!(n-k)!}\,.$$

Проверим, что вероятности p_k соответствуют закону распределения:

$$\sum_{k=1}^{4} p_k = \frac{5^4 + 4 \cdot 5^3 + 6 \cdot 5^2 + 4 \cdot 5 + 1}{6^4} = \frac{(5+1)^4}{6^4} = 1.$$

число успехов k	варианты комбинаций	вероятность p_k
0	○ ○ ○○	$C_4^0 p^0 q^4 = \left(\frac{5}{6}\right)^4$
1	● ○ ○○ ○ ● ○○ ○ ○ ●○ ○ ○ ○●	$C_4^1 p^1 q^3 = 4 \cdot \frac{1}{6}\left(\frac{5}{6}\right)^3$
2	● ● ○○ ● ○ ●○ ● ○ ○● ○ ● ●○ ○ ● ○● ○ ○ ●●	$C_4^2 p^2 q^2 = 6 \cdot \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^2$
3	● ● ●○ ● ● ○● ● ○ ●● ○ ● ●●	$C_4^3 p^3 q^1 = 4 \cdot \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^1$
4	● ● ●●	$C_4^4 p^4 q^0 = \left(\frac{1}{6}\right)^4$

Таблица 1. Биномиальный закон распределения при $n = 4$ и $p = 0.5$.

Обозначим через ε_i, $i = 1, 2, \ldots, n$, независимые случайные величины, имеющие распределение Бернулли с параметром p. Случайная величина $\xi = \varepsilon_1 + \ldots + \varepsilon_n$, равная случайному числу успехов из n независимых испытаний, имеет **биномиальный закон распределения**:

$$P\{\xi = k\} = C_n^k p^k (1-p)^{n-k},\ k = 0, 1, \ldots, n. \tag{9}$$

Для краткости можно использовать обозначение $\xi \sim \beta(p, n)$.

Отсюда получаем, что дисперсия биномиально распределенной случайной величины равна npq. Производящая функция моментов, коэффициенты асимметрии и эксцесса представлены в таблице 2.

В R для моделирования n биномиальных случайных величин с вероятностью успеха $prob$ из $size$ независимых испытаний используется функция $rbinom(n, size, prob)$. Для моделирования случайных величин с распределением Бернулли нужно указать $size = 1$.

4.2. Распределение Пуассона

Отдельно рассматривается случай, когда число испытаний велико при малой вероятности успеха p. Например, вероятность вызова скорой помощи в определенном временном интервале чрезвычайно мала в каждом отдельном случае. Но в большом городе при наличии большого числа испытаний количество вызовов может оказаться счетным, то есть равным $0, 1, 2, \ldots$. Для того чтобы вычислить при $p \to 0$ и $n \to \infty$ предел вероятности

$$p_k = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k},$$

обозначим через $\lambda = np$ и умножим и разделим p_k на n^k.

$$p_k = \frac{n(n-1)\ldots(n-k+1)}{k! n^k} (np)^k (1-p)^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda}, \tag{10}$$

так как, по замечательному пределу, $e^{-1} = \lim\limits_{p \to 0} (1-p)^{\frac{1}{p}}$, при заданном k

$$\lim_{p \to 0} (1-p)^{n-k} = \lim_{p \to 0} \left((1-p)^{\frac{1}{p}} \right)^{pn} (1-p)^{-k} = e^{-\lambda},$$

$$\lim_{n \to \infty} \frac{n(n-1)\ldots(n-k+1)}{n^k} = 1\,.$$

Случайная величина ξ, принимающая значения $k = 0, 1, 2, \ldots$ с вероятностями (10), имеет распределение Пуассона $\mathcal{P}(\lambda)$. Так как $e^{\lambda} = \sum\limits_{k=0}^{\infty} \frac{\lambda^k}{k!}$, то

$$\sum_{k=0}^{\infty} p_k = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = 1\,.$$

Производящая функция моментов имеет вид

$$g(\nu) = \sum_j \frac{\lambda^j}{j!} e^{-\lambda} \nu^j = e^{-\lambda} \sum_j \frac{(\lambda\nu)^j}{j!} = e^{-\lambda+\lambda\nu}\,. \tag{11}$$

Математическое ожидание и дисперсия случайной величины ξ равны λ, так как

$$\mathbb{E}\xi = \sum_{k=0}^{\infty} k p_k = \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} = \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} = \lambda\,.$$

$$\mathbb{E}\xi^2 = \sum_{k=0}^{\infty} k^2 p_k = \sum_{k=1}^{\infty} \frac{k\lambda^k}{(k-1)!} e^{-\lambda} \overset{k=t+1}{=} \sum_{t=0}^{\infty} \frac{(t+1)\lambda^{t+1}}{t!} e^{-\lambda} = \lambda^2 + \lambda,$$

$$\mathbb{D}\xi = \mathbb{E}\xi^2 - (\mathbb{E}\xi)^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda\,.$$

Проверим правильность полученных выражений при помощи производящих функций из (2):

$$g'(\nu) = (e^{-\lambda+\lambda\nu})' = \lambda e^{-\lambda+\lambda\nu}|_{\nu=1} = \lambda,$$

$$g''(\nu) = \lambda^2 e^{-\lambda+\lambda\nu}|_{\nu=1} = \lambda^2,$$

$$\mathbb{D}\xi = g''(1) + g'(1) - (g'(1))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda\,.$$

Коэффициенты асимметрии и эксцесса представлены в таблице 2. В R для моделирования n случайных величин, распределенных по закону Пуассона с интенсивностью λ используется функция $rpois(n, \lambda)$.

Покажем, что коэффициент асимметрии и эксцесс распределения Пуассона $\mathcal{P}(\lambda)$ равны соответственно $\frac{1}{\sqrt{\lambda}}$ и $\frac{1}{\lambda}$.

Рассмотрим выражения для начальных моментов $\alpha_k = \mathbb{E}\xi^k$.

$$\alpha_k = \mathbb{E}\xi^k = \sum_{j=1}^{\infty} \frac{j^{k-1}\lambda^j}{(j-1)!} e^{-\lambda} \quad \overset{j=t+1}{=} \lambda \sum_{t=0}^{\infty} \frac{(t+1)^{k-1}\lambda^t}{t!} e^{-\lambda}.$$

$$\alpha_1 = \lambda, \quad \alpha_2 = \lambda^2 + \lambda,$$

$$\alpha_3 = \lambda(\alpha_2 + 2\alpha_1 + 1) = \lambda(\lambda^2 + \lambda + 2\lambda + 1) = \lambda^3 + 3\lambda^2 + \lambda,$$

$$\alpha_4 = \lambda(\alpha_3 + 3\alpha_2 + 3\alpha_1 + 1) =$$

$$= \lambda(\lambda^3 + 3\lambda^2 + \lambda + 3\lambda^2 + 3\lambda + 3\lambda + 1) = \lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda,$$

Вычислим центральные моменты $\mu_k = \mathbb{E}(\xi - \mathbb{E}\xi)^k$ при $k = 3, 4$.

$$\mu_3 = \alpha_3 - 3\alpha_2\alpha_1 + 2\alpha_1^3 = (\lambda^3 + 3\lambda^2 + \lambda) - 3\lambda(\lambda^2 + \lambda) + 2\lambda^3 = \lambda\,,$$

$$\mu_4 = \alpha_4 - 4\alpha_3\alpha_1 + 6\alpha_2\alpha_1^2 - 3\alpha_1^4 =$$

$$= (\lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda) - 4\lambda(\lambda^3 + 3\lambda^2 + \lambda) + 6\lambda^2(\lambda^2 + \lambda) - 3\lambda^4 =$$

$$= 3\lambda^2 + \lambda\,,$$

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\lambda}{\lambda^{3/2}} = \frac{1}{\sqrt{\lambda}},\quad \gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{3\lambda^2 + \lambda}{\lambda^2} - 3 = \lambda^{-1}\,.$$

4.3. Равномерное распределение.

Случайная величина $\xi \sim U(a, b)$ равномерно распределена на интервале $[a, b]$, если ее плотность распределения имеет вид $f(x) = \frac{1}{b-a}$ на интервале $[a, b]$ и равна нулю вне этого интервала. МО равно середине интервала, дисперсия пропорциональна длине интервала.

$$\mathbb{E}\xi = \int_a^b \frac{x}{b-a}dx = \frac{1}{b-a}\frac{x^2}{2}\Big|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2},$$

$$\mathbb{E}\xi^2 = \int_a^b \frac{x^2}{b-a}dx = \frac{1}{b-a}\frac{x^3}{3}\Big|_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3},$$

$$\mathbb{D}\xi = \mathbb{E}\xi^2 - (\mathbb{E}\xi)^2 = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}\,.$$

Коэффициенты асимметрии и эксцесса представлены в таблице 2. В R для моделирования n случайных величин, распределенных по равномерному закону на интервале $[a, b]$ используется функция $runif(n, min = a, max = b)$. Покажем, что эксцесс равномерного распределения равен $-\frac{6}{5}$. Так как эксцесс не зависит от сдвига и масштаба, рассмотрим случай $\xi \sim U(0, 1)$.

$$\alpha_k = \int\limits_0^1 x^k dx = \frac{1}{k+1}\,.$$

$$\mu_2 = \alpha_2 - \alpha_1^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12},$$

$$\mu_3 = 0,$$

$$\mu_4 = \alpha_4 - 4\alpha_3\alpha_1 + 6\alpha_2\alpha_1^2 - 3\alpha_1^4 = \frac{1}{5} - \frac{4}{4\cdot 2} + \frac{6}{3\cdot 4} - \frac{3}{16} = \frac{1}{80},$$

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{12^2}{80} - 3 = -\frac{96}{80} = -\frac{6}{5}\,.$$

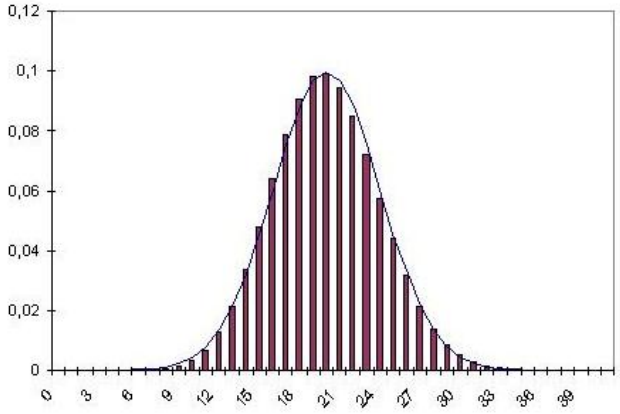

Рис. 2. Биномиальное распределение с параметрами $n = 100$ и $p = 0.2$ и нормальное распределение с параметрами $\mu = 20$, $\sigma^2 = 16$.

4.4. Нормальное распределение

Нормальное или гауссовское распределение $\mathcal{N}(\mu, \sigma)$ было получено как предельное биномиальное распределение при увеличении числа испытаний. Плотность его распределения зависит от двух параметров μ, σ^2, смысл которых заключается соответственно в среднем и дисперсии.

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{12}$$

При $\mu = 0$, $\sigma^2 = 1$ распределение $f(x|0, 1)$ называется стандартным нормальным.

Для доказательства основного свойства плотности нужно вычислить двойной интеграл

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}-\frac{y^2}{2}}dxdy = 4\int_{0}^{\infty}\int_{0}^{\infty} e^{-\frac{x^2}{2}-\frac{y^2}{2}}dxdy = \\ = \left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}dx\right)\left(\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}dy\right),$$

который может быть вычислен переходом к полярным координатам $x = r\cos\phi, y = r\sin\phi$ с якобианом преобразования

$$dxdy = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \phi} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \phi} \end{vmatrix} drd\phi = \begin{vmatrix} \cos\phi & -r\sin\phi \\ \sin\phi & r\cos\phi \end{vmatrix} drd\phi = rdrd\phi,$$

$$4\int_{0}^{\infty}\int_{0}^{\infty} e^{-\frac{x^2}{2}-\frac{y^2}{2}}dxdy = 4\int_{0}^{\frac{\pi}{2}} d\phi \int_{0}^{\infty} e^{-\frac{r^2}{2}}rdr = 2\pi\left(-e^{-\frac{r^2}{2}}\right)|_0^{\infty} = 2\pi\,,$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}-\frac{y^2}{2}}dxdy = \left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}dx\right)^2 = 2\pi,$$

И основное свойство плотности справедливо за счет

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}\,. \tag{13}$$

Пусть $\xi_0 \sim \mathcal{N}(0,1)$. Математическое ожидание равно нулю $\mathbb{E}\xi_0 = 0$ из-за симметричности функции относительно нуля:

$$\mathbb{E}\xi_0 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x e^{-\frac{x^2}{2}} dx = 0\,,$$
$$\mathbb{E}\xi_0^2 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x^2 e^{-\frac{x^2}{2}} dx = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} x \cdot x e^{-\frac{x^2}{2}} dx =$$

Воспользуемся интегрированием по частям $\int uv'dx = uv - \int u'vdx$, где $u = x$, $v' = xe^{-\frac{x^2}{2}}$, $v = -e^{-\frac{x^2}{2}}$.

$$= -\frac{2}{\sqrt{2\pi}} x e^{-\frac{x^2}{2}} |_0^{\infty} + \frac{2}{\sqrt{2\pi}} \int_0^{\infty} e^{-\frac{x^2}{2}} dx = 1\,.$$

Следовательно, $\mathbb{D}\xi_0 = 1$. Отсюда для $\xi = \sigma\xi_0 + \mu$ имеем математическое ожидание $\mathbb{E}\xi = \mathbb{E}(\sigma\xi_0 + \mu) = \mu$ и дисперсию $\mathbb{D}\xi = \mathbb{D}(\sigma\xi_0 + \mu) = \sigma^2\mathbb{D}\xi_0 = \sigma^2$.

Коэффициенты асимметрии и эксцесса представлены в таблице 2. В R для моделирования n случайных величин, распределенных по нормальному закону с параметрами μ, σ, используется функция $rnorm(n, mean = \mu, sd = \sigma)$.

4.5. Гамма распределение

Прежде всего нужно отметить важную роль гамма распределения в математической статистике. Его существенный частный случай виде распределения хи-квадрат используется для большинства статистических критериев. Для того чтобы этот факт не остался абстратным, рассмотрим это распределение более подробно. Свое название этот закон получил из-за того, что в выражение для его плотности входит гамма-функция, которой 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.

$$\Gamma(\lambda+1)=\int\limits_0^\infty \underbrace{x^\lambda}_{u}\underbrace{e^{-x}}_{v'}dx=uv\,|_0^\infty-\int\limits_0^\infty u'vdx=$$

$$=x^\lambda\left(-e^{-x}\right)|_0^\infty-\int\limits_0^\infty\lambda x^{\lambda-1}\left(-e^{-x}\right)dx=\lambda\Gamma(\lambda)\,,$$

$$\Gamma(1)=\int\limits_0^\infty x^0e^{-x}dx=-e^{-x}\,|_0^\infty=1\,.$$

Если $\lambda\in\mathbf{N}$, то повторное применение $\Gamma(n+1)=n\Gamma(n)$ приведет к $\Gamma(n+1)=n!$. Последнее получаем из равенства (13), то есть из $\int\limits_{-\infty}^{\infty}e^{-\frac{t^2}{2}}dt=\sqrt{2\pi}$. Используя замену $\sqrt{x}=\frac{t}{\sqrt{2}},\ \ x=\frac{t^2}{2},\ \ dx=tdt$, получаем

$$\Gamma\left(\frac{1}{2}\right)=\int\limits_0^\infty\frac{1}{\sqrt{x}}e^{-x}dx=\int\limits_0^\infty\frac{\sqrt{2}}{t}e^{-\frac{t^2}{2}}tdt=\sqrt{2}\frac{\sqrt{2\pi}}{2}=\sqrt{\pi}\,.$$

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

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

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

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 $\xi_1 + \xi_2 \sim \gamma(x, \alpha, \lambda_1 + \lambda_2)$, то есть при одинаковом параметре масштаба параметры формы просто складываюися.

Для вычисления первых двух моментов можно воспользоваться свойством характеристической функции $\psi^{(\nu)}(0) = i^\nu \alpha_\nu$.

$$\begin{aligned}
\psi'(t) &= -\lambda \left(1 - \frac{it}{\alpha}\right)^{-\lambda-1} \left(-\frac{i}{\alpha}\right), \\
\psi''(t) &= (-\lambda)(-\lambda-1) \left(1 - \frac{it}{\alpha}\right)^{-\lambda-2} \left(-\frac{i}{\alpha}\right)^2, \\
\alpha_1 &= \frac{\psi'(0)}{i} = \frac{\lambda}{\alpha}, \\
\alpha_2 &= \frac{\psi''(0)}{i^2} = \frac{1}{i^2}\lambda(\lambda+1)\left(-\frac{i}{\alpha}\right)^2 = \frac{\lambda^2+\lambda}{\alpha^2}, \\
\mu_2 &= \alpha_2 - \alpha_1^2 = \frac{\lambda}{\alpha^2}.
\end{aligned}$$

4.6. 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 $f_\eta(x)$ рас-

пределения 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 $\xi_1^2 + \ldots + \xi_n^2$, 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:

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

4.7. Распределение Вейбулла

Пусть случайная величина ξ имеет распределение Вейбулла $W(k, \lambda)$, которое задается плотностью

$$w(x|k,\lambda) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k},\ x \ge 0.$$

Покажем, что случайная величина $\eta = \xi^k$ имеет гамма распределением с параметром формы $\lambda = 1$ и масштаба $\alpha = \lambda^k$, то есть экспоненциальное

распределение. Действительно,

$$F_\eta(x) = \mathbb{P}\{\eta \le x\} = \mathbb{P}\{\xi^k \le x\} = \mathbb{P}\{\xi \le x^{\frac{1}{k}}\} = F_\xi(x^{\frac{1}{k}}),$$
$$f_\eta(x) = F'_\eta(x) = F'_\xi(x^{\frac{1}{k}}) \cdot \frac{1}{k} x^{\frac{1}{k}-1} =$$
$$= \frac{k}{\lambda}\left(\frac{x^{\frac{1}{k}}}{\lambda}\right)^{k-1} e^{-\left(\frac{x^{\frac{1}{k}}}{\lambda}\right)^k} \cdot \frac{1}{k} x^{\frac{1}{k}-1} = \frac{1}{\lambda^k} e^{-\frac{x}{\lambda^k}}.$$

Числовые характеристики этого распределения указаны в табл.2 и в замечаниях к ней. Например, для вычисления математического ожидания делаем замену $t = \left(\frac{x}{\lambda}\right)^k$, $x = \lambda t^{\frac{1}{k}}$, $dx = \frac{\lambda}{k} t^{\frac{1}{k}-1}$,

$$\mathbb{E}\xi = \int_0^\infty \frac{kx}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k} dx =$$
$$= \int_0^\infty kte^{-t}\frac{\lambda}{k} t^{\frac{1}{k}-1} dt = \lambda \int_0^\infty t^{\frac{1}{k}} e^{-t} dt = \lambda\Gamma\left(\frac{1}{k}+1\right).$$

$$\mathbb{E}\xi^2 = \int_0^\infty \frac{kx^2}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k} dx =$$
$$= \int_0^\infty k(\lambda t^{\frac{1}{k}})te^{-t}\frac{\lambda}{k} t^{\frac{1}{k}-1} dt = \lambda \int_0^\infty t^{\frac{1}{k}} e^{-t} dt = \lambda^2\Gamma\left(\frac{2}{k}+1\right),$$
$$\text{отсюда } \mathbb{D}\xi = \lambda^2(\Gamma(1+2k^{-1}) - \Gamma^2(1+k^{-1})).$$

4.8. Обобщенное Пуассоновское распределение

Пусть $\{\xi_k\}$ последовательность взаимно независимых случайных величин с распределением $\mathbb{P}\{\xi_k = j\} = f_j$ и производящей функцией $f(\nu) = \sum f_j \nu^j$. Рассмотрим случайную сумму случайных величин $\zeta_\tau = \xi_1 + \ldots + \xi_\tau$, где τ имеет распределение $\mathbb{P}\{\tau = n\} = g_n$ с производящей функцией $g(\nu) = \sum g_j \nu^j$ и не зависит от компонент ξ_j. Распределение ζ_τ можно вычислить по формуле полной вероятности

$$h_j = \mathbb{P}\{\zeta_\tau = j\} = \sum_{n=0}^\infty \mathbb{P}\{\tau = n\}\mathbb{P}\{\xi_1 + \ldots + \xi_n = j\}.$$

Покажем, что производящая функция $h(\nu) = \sum h_j \nu^j = g(f(\nu))$. По определению производящей функции и свойству их произведения в случае неза-

висимых величин $\mathbb{E}(\nu^{\zeta_\tau}|\tau = n) = f^n(\nu)$. Отсюда

$$h(\nu) = \mathbb{E}(\nu^{\zeta_\tau}) = \sum_n \mathbb{P}\{\tau = n\}\mathbb{E}(\nu^{\zeta_\tau}|\tau = n) = \sum_n \mathbb{P}\{\tau = n\} f^n(\nu) = g(f(\nu)).$$

Если τ имеет распределение Пуассона $g_j = \frac{\lambda^j}{j!}e^{-\lambda}$, $j = 0, 1, 2, \ldots$, с производящей функцией $g(\nu) = e^{-\lambda+\lambda\nu}$ из (11), то говорят, что ζ_τ имеет обобщенное распределение Пуассона. Пусть внутреннее распределение f_j задается математическим ожиданием μ и дисперсией σ^2. Получим выражения для первых двух моментов для случайной величины ζ_τ.

$$h(\nu) = e^{-\lambda+\lambda f(\nu)},\ h'(\nu) = e^{-\lambda+\lambda f(\nu)}\lambda f'(\nu)\,,\quad h'(1) = \lambda\mu$$
$$h''(\nu) = e^{-\lambda+\lambda f(\nu)}(\lambda f'(\nu))^2 + e^{-\lambda+\lambda f(\nu)}\lambda f''(\nu)\,,$$
$$h''(1) = \lambda^2\mu^2 + \lambda(\mu^2 + \sigma^2),$$
$$\mathbb{E}\zeta_\tau = \lambda\mu,\quad \mathbb{D}\zeta_\tau = \lambda(\mu^2 + \sigma^2)\,.$$

В частности, если $\tau \sim \mathcal{P}(\lambda_1)$, $\xi_j \sim \mathcal{P}(\lambda_0)$, то $\mathbb{E}\zeta_\tau = \lambda_1\lambda_0$, $\mathbb{D}\zeta_\tau = \lambda_1(\lambda_0^2 + \lambda_0)$.

5. Методы оценки параметров

Если функция распределения известна с точностью до параметров $\theta_1, \ldots, \theta_r$, например, как функция гамма распределения с параметрами α и λ, то для оценки этих параметров используют методы моментов и максимального правдоподобия.

Метод моментов состоит в приравнивании эмпирических и теоретических выражений для параметров и в решении соответствующей системы уравнений. Например, для модели гамма распределения с параметрами α и λ получены выборочные оценки среднего и дисперсии, соответственно $\bar{x}$

и S^2 из (6). Для оценки параметров решается система

$$\begin{cases} \frac{\lambda}{\alpha} = \bar{x}, \\ \frac{\lambda}{\alpha^2} = S^2, \end{cases}$$

из которой получаем оценки $\hat{\alpha} = \frac{\bar{x}}{S^2}, \quad \hat{\lambda} = \frac{\bar{x}^2}{S^2}$.

Метод максимального правдоподобия предполагает в качестве оценок параметров $\Theta = (\theta_1, \ldots, \theta_r)$ такие значения, при которых максимальна функция правдоподобия $\mathcal{L}(x_1, \ldots, x_n|\Theta)$, то есть вероятность появления данной выборки. В случае дискретных распределений функция распределения представляет собой произведение вероятностей

$$\mathcal{L}(x_1, \ldots, x_n|\Theta) = \prod_{i=1}^{n} \mathbb{P}_\Theta\{\xi = x_i\},$$

а в случае непрерывного — произведение плотностей

$$\mathcal{L}(x_1, \ldots, x_n|\Theta) = \prod_{i=1}^{n} f_\Theta(x_i).$$

При решении стандартных задач можно воспользоваться встроенными функциями, например, mle, пакет $stats4$ в R. В нестандартных ситуациях можно решить задачу численной оптимизации при помощи $optim$ или какой-то другой функции.

Итак, для моделирования выборки объема N применяются следующие функции:

- $rbinom(N, size, prob)$ для биномиального $\beta(n, p)$, где $size$ означает число независимых испытаний n, $prob$ означает вероятность p успеха в одном испытании;
- $rpois(N, lambda)$ для распределения Пуассона, где $lambda$ соответствует параметру интенсивности λ;

Распределение	Среднее	Дисперсия	γ_1	γ_2
$\beta(p,n)$	np	npq	$\frac{q-p}{\sqrt{npq}}$	$\frac{1-6qp}{npq}$
$\mathcal{P}(\lambda)$	λ	λ	$\lambda^{-\frac{1}{2}}$	λ^{-1}
$\beta_-(k,p)$	$\frac{k(1-p)}{p}$	$\frac{k(1-p)}{p^2}$	$\frac{2-p}{\sqrt{k(1-p)}}$	$\frac{6}{k}+\frac{p^2}{k(1-p)}$
$\gamma(\lambda,\alpha)$	$\frac{\lambda}{\alpha}$	$\frac{\lambda}{\alpha^2}$	$\frac{2}{\sqrt{\lambda}}$	$\frac{6}{\lambda}$
$W(k,\lambda)$	$\lambda\Gamma(1+k^{-1})$	$\lambda^2(\Gamma(1+2k^{-1})-\Gamma^2(1+k^{-1})$		
$U(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	0	$-\frac{6}{5}$
$\mathcal{N}(\mu,\sigma)$	μ	σ^2	0	0

Таблица 2. Характеристики распределений. Для распределения Вейбулла асимметрия и эксцесс имеют соответственно вид: $\gamma_1 = \frac{\lambda^3\Gamma(1+3k^{-1})-3\mu\lambda^2\Gamma(1+2k^{-1})+2\mu^2}{\sigma^3}$, $\gamma_2 = \frac{\lambda^4\Gamma(1+4k^{-1})-4\lambda^3\mu\Gamma(1+3k^{-1})+6\mu^2\lambda^2\Gamma(1+2k^{-1})-3\mu^4}{\sigma^4}$.

- $rnbinom(N, size, prob)$ для отрицательно биномиального $\beta_-(k,p)$, где $size$ и $prob$ соответствуют параметрам k и p.

- $rgamma(N, shape, rate = 1, scale = 1/rate)$ для гамма распределения $\Gamma(\alpha,\lambda)$ с плотностью $\gamma(x|\alpha,\lambda) = \frac{\alpha^\lambda}{\Gamma(\lambda)}x^{\lambda-1}e^{-\alpha x}$, $x \geq 0$, где $shape$ и $rate$ соответствуют параметрам λ и α.

- $rweibull(N, shape, scale)$ для распределения Вейбулла $W(k,\lambda)$, где $shape$ соответствует параметру k, а $scale$ параметру λ;

- $runif(N, min, max)$ для равномерного $U(a,b)$, где min, max соответствуют параметрам a, b.

- для нормального $\mathcal{N}(\mu,\sigma)$, где $mean$, sd соответствуют μ, σ.

6. Проверка согласия эмпирического распределения с теоретическим

На рис. 3 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 $\hat{\mu} = \bar{x} = 48.72$, $\hat{\sigma} = S = 20.47$.

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[1].

$$p_i = \Phi\left(\frac{z_i - \hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{z_{i-1} - \hat{\mu}}{\hat{\sigma}}\right). \tag{14}$$

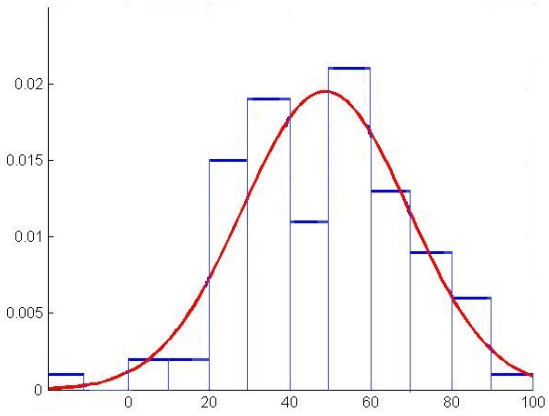

Рис. 3. Выборочная гистограмма и плотность $\mathcal{N}(\mu, \sigma)$.

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

[1] В электронных таблицах $\Phi(x)$ можно получить при помощи функции НОРМСТРАСП(x), в R функция $pnorm(x)$.

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 распределена по определенному закону, в множестве значений случайной величины выделяется критическую область, вероятность α попадания в которую настолько мала, что мы считаем это невозможным. Если случайная величина туда все-таки попала, то говорят, что она скорее имеет какое-то другое распределение. Если случайная величина не попала в критическую область, то говорят, что оснований отвергнуть гипотезу нет. Вероятность α попадания в критическую область называется уровнем значимости.

Согласно теореме Пирсона, статистика $\chi^2 = \sum\limits_{i=1}^{r} \frac{(\nu_i-np_i)^2}{np_i}$ при условии, что выборка подчиняется данному закону распределения, имеет распределение хи-квадрат с $r-1$ степенью свободы. Обобщение данного критерия позволяет проверять гипотезы согласия в случае, когда параметры распределения оцениваются по выборке. В таком случае число степеней свободы уменьшается на число оцениваемых параметров.

Критическое значение, соответствующее $\alpha = 0.05$, вычисляется при помощи таблиц, калькуляторов, *Excel*, *R* и равно $\chi^2_{0.95} = \text{ХИ2ОБР}(0.05; 5) =$

11.07.[2] Наблюдаемое значение статистики $\chi^2_* = 8.356 < \chi^2_{0.95}$, следовательно, гипотеза о согласии эмпирического распределения с нормальным не отвергается с уровнем значимости $\alpha = 0.05$. Аналогичный вывод можно сделать при помощи доверительного уровня вероятности или $p-$значения [3] $p = P\{\chi^2 > \chi^2_*\} = \text{ХИ2РАСП}(8.356; 5) = 0.14 > \alpha$, то есть нет оснований отвергнуть гипотезу, которая отвергается только при $p < \alpha$. Для проверки нормальности вместо критерия Пирсона можно использовать специальный критерий Шапиро-Уилка, *shapiro.test*() в *R*.

7. Варианты практического задания

1. Промоделировать распределение

   - биномиальное $\beta(n,p)$, $\mathbb{P}\{\xi = k\} = C_n^k p^k (1-p)^{n-k},\ k = 0, 1, \ldots, n.$
   - Пуассона $P(\lambda)$, $\mathbb{P}\{\xi = j\} = \frac{\lambda^j}{j!} e^{-\lambda},\ j = 0, 1, \ldots, \infty.$
   - отрицательно биномиальное $\beta_-(k,p)$, $\mathbb{P}\{\xi = j\} = \frac{\Gamma(k+j)}{\Gamma(k)j!} p^k (1-p)^j,\ j = 0, 1, \ldots, \infty.$
   - гамма $\Gamma(\alpha, \lambda)$ с плотностью $\gamma(x|\alpha,\lambda) = \frac{\alpha^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\alpha x},\ x \geq 0.$
   - Вейбулла $W(k,\lambda)$ с плотностью $w(x|k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k},\ x \geq 0.$
   - равномерное $U(a,b)$ с плотностью $u(x|a,b) = \frac{1}{b-a},\ a \leq x \leq b.$
   - нормальное $\mathcal{N}(\mu,\sigma)$ с плотностью $f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$
   - обобщенное распределение Пуассона с внутренним также распределением Пуассона и с соответствующими параметрами λ_1, λ_2.

[2] В *R* аналогичная функция $qchi(0.95, 5)$.

[3] В *R* используется $pchisq(x, df)$

2. На модельных данных построить гистограмму, найти математическое ожидание, дисперсию, стандартное отклонение, ошибку среднего, медиану, минимум, максимум, квартили и интервартильный размах, асимметирю, эксцесс. Сравнить выборочные характеристики с теоретическими.

3. Оценить параметры распределения по методу моментов и по методу максимального правдоподобия.

4. Проверить согласие эмпирического и теоретического распределения по критерию хи-квадрат Пирсона.

8. Пример выполнения расчетов в R

На рис.4 представлен фрагмент отчета по моделированию нормального закона распределения и вычисления характеристик модельной выборки. Библиотека moments нужна для проверки вычисления асимметрии и эксцесса. Результаты счета сопровождаются сравнением полученных оценок с параметрами распределения. На рис.5 приведен код и результаты вычисления медианы: при помощи встроенной функции и непосредственно по правилу. Убеждаемся, что результаты совпадают. На рис.6 показно, как двумя способами можно вычислить квартили: при помощи функций $quantile()$ и $pnorm()$. Вычисление асимметрии и эксцесса показано на рис.7. Также применется два способа вычисления: по формулам и при помощи встроенной функции. Небольшое отклонение от нуля в отрицательную сторону свидетельствует о несколько большей концентрации наблюдений справа и о левом хвосте. Эксцесс также немного отличается от нуля в отрицательную сторону, что свидетельствует о чуть меньшей выраженности эксцесса и более сглаженном пике.

```
library(moments)
#параметры нормального распределения
mu<-20; sigma<-5;N<-200
#моделирование
X<-rnorm(N,mean=mu,sd=sigma)

#математическое ожидание
mu.<-mean(X);
c(mu.=mu.,mu=mu)

##      mu.       mu
## 19.75541 20.00000

#стандартное отклонение
sigma.<-sd(X)
c(sigma.=sigma.,sigma=sigma)

##   sigma.    sigma
## 5.436528 5.000000

#дисперсия
sigma2.<-sigma.^2
c(sigma2.=sigma2.,sigma2=sigma^2)

##  sigma2.   sigma2
## 29.55584 25.00000

#ошибка среднего
c(err.=sigma./sqrt(N),err=sigma/sqrt(N))

##      err.       err
## 0.3844206 0.3535534
```

Рис. 4. Моделирование, вычисление среднего и характеристик вариабельности выборки.

```
#медиана
X1<-sort(X)
c(median.=median(X),Median=mean(c(X1[N/2],X1[N/2+1])))

##  median.   Median
## 19.67056 19.67056

#минимум, максимум, размах
c(Min=min(X),Max=max(X), R=max(X)-min(X))

##       Min       Max         R
##  4.567321 32.669389 28.102068
```

Рис. 5. Вычисление медианы и размаха выборки.

```
#квартили,  и Q интервартильный размах при помощи функции  quantile
q<-quantile(X, probs = seq(0, 1, 0.25) );q

##        0%       25%       50%       75%      100%
##  4.567321 16.552610 19.670560 23.705324 32.669389

Q<-q[4]-q[2];Q

##      75%
## 7.152714

#генеральные  квартили,  и Q интервартильный размах при помощи функции
qnorm
q<-c(qnorm(0.25,mean=mu.,sd=sigma),qnorm(0.75,mean=mu.,sd=sigma));q

## [1] 16.62755 23.37245

q[2]-q[1]

## [1] 6.744898

#эмпирические  квартили,  и Q интервартильный размах
q.<-
c(qnorm(0.25,mean=mu.,sd=sigma.),qnorm(0.75,mean=mu.,sd=sigma.));q.

## [1] 16.08853 23.42230

q.[2]-q.[1]

## [1] 7.333765
```

Рис. 6. Вычисление интервартильного размаха выборки.

```
#асимметрия
c(mean((X-mu.)^3),m3<-moment(X, order = 3, central = TRUE))

## [1] -31.54361 -31.54361

m2<-moment(X, order = 2, central = TRUE)
c(m3/m2^(3/2),skewness(X))

## [1] -0.1977935 -0.1977935

#эксцесс
c(mean((X-mean(X))^4),m4<-moment(X, order = 4, central = TRUE))

## [1] 2582.642 2582.642

c(kurtosis(X)-3,m4/m2^2-3)

## [1] -0.01371237 -0.01371237
```

Рис. 7. Вычисление асимметрии и эксцесса.

```
#Оценка параметров
f<-function(x)dnorm(x,mean=mu.,sd=sigma)
f.<-function(x)dnorm(x,mean=mu.,sd=sigma.)

#Метод максимального правдоподобия
# функция правдоподобия
Func.prob.log <- function(x) -sum(dnorm(X, mean = x[1], sd = x[2], log
= TRUE))


#Оценки ММП
res<-optim(c(mu.,sigma.),Func.prob.log)
mu..<-res$par[1]; sigma..<-res$par[2];
c(mu..=mu..,sigma..=sigma..)

##      mu..   sigma..
## 19.755967  5.423307
```

Рис. 8. Оценки по методу максимального правдоподобия.

Поскольку для нормального закона распределения параметрами являются математическое ожидание и дисперсия, система для оценки параметров по методу моментов (ММ) оказывается тривиальной, и оценки ММ представлены на рис. 4. Для получения оценок по методу максимального правдоподобия нужно написать функцию правдоподобия (рис.8) и решить экстремальную задачу, то есть численно найти значения, при которых логарифм этой функции будет максимальным. В данном случае применена функция *optim*(). На рис.9 и 10 приведен пример кода для построения графиков и сами графики, иллюстрирующие согласосанность эмпирического и теоретического распределений. Плотности распределений, построенные по оценкам параметров, оказались близкими, но отличными от плотности, построенной по заданным значениям параметров.

При построении критерия согласия Пирсона (рис.11) нужно быть уверенным в том, что вероятности попадания в интервалы вычислены верно, и что их сумма равна 1.

Таблица соответствия между эмпирическими и теоретическими часто-

```
f..<-function(x)dnorm(x,mean=res$par[1],sd=res$par[2])
hist(X,freq=FALSE,xlim=c(min(X),max(X)),main="")
curve(f,min(X),max(X),add=TRUE,col=2)
curve(f.,min(X),max(X),add=TRUE,col=3)
curve(f..,min(X),max(X),add=TRUE,col=4,lty=2)
legend('topright',c("hyp","mm","mmp"),pch=20,col=c(2,3,4),lty=c(1,1,2)
)
```

Рис. 9. Код для построения графиков, иллюстрирующих согласие эмпирического и теоретического распределения.

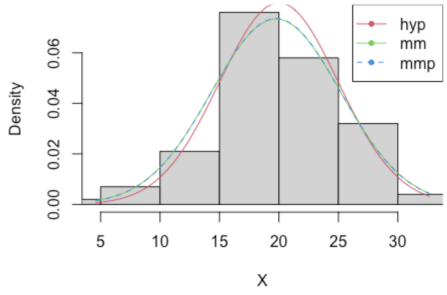

Рис. 10. Гистограмма эмпирического распределения и кривые плотности нормального закона с параметрами: hyp - заданными, mm - оцененными по ММ, mmp - по ММП.

```
#Проверка гипотезы согласия
h<-hist(X,plot=FALSE)

#эмпирические частоты
n.i<-sapply(seq(length(h$breaks)-1)+1,function(i)
length(X[X<h$breaks[i] & X>=h$breaks[i-1]]))
#гипотетические частоты
p.i<-sapply(seq(length(h$breaks)-1)+1,function(i)
    pnorm(h$breaks[i],mean=mu..,sd=sigma..)-
pnorm(h$breaks[i-1],mean=mu..,sd=sigma..))
sum(p.i)

## [1] 0.9973945

p.i[1]<-pnorm(h$breaks[2],mean=mu..,sd=sigma..)
p.i[length(p.i)]<-1-
pnorm(h$breaks[length(h$breaks)-1],mean=mu..,sd=sigma..)
sum(p.i)

## [1] 1
```

Рис. 11. Построение критерия согласия Пирсона.

```
#-------------
#проверка условия n*p_i>5
tab<-cbind(h$counts,p_i*length(X));tab

##      [,1]       [,2]
## [1,]    2  0.6511568
## [2,]    7  6.5523515
## [3,]   21 30.8478899
## [4,]   76 65.5376318
## [5,]   58 63.0537500
## [6,]   32 27.4665672
## [7,]    4  5.8906527

t1<-cumsum(tab[,2]);T1<-min(which(t1>5));T1

## [1] 2

t2<-cumsum(tab[seq(nrow(tab),1,-1),2]);t2;T2<-min(which(t1>5));T2<-
nrow(tab)-T2;T2

## [1]   5.890653  33.357220  96.410970 161.948602 192.796492
199.348843 200.000000

## [1] 5

Tab<-apply(tab,2,function(x)
c(sum(x[seq(T1)]),x[c((T1+1):T2)],sum(x[c((T2+1):nrow(tab))])))
Tab

##      [,1]      [,2]
## [1,]    9  7.203508
## [2,]   21 30.847890
## [3,]   76 65.537632
## [4,]   58 63.053750
## [5,]   36 33.357220
```

Рис. 12. Проверка условия $np_i > 5$.

```
#статистика хи-квадрат
chi2<-sum(apply(Tab,1,function(x) (x[1]-x[2])^2/x[2]))
#доверительный уровень вероятности
p.value<-1-pchisq(chi2,nrow(Tab)-3);
print(paste("p.value",round(p.value,4),sep="="))

## [1] "p.value=0.053"
```

Рис. 13. Вычисление статистики хи-квадрат и доверительного уровня вероятностей.

тами должна быть построена таким образом, чтобы для каждой i-й ячейки выполнялось условие $np_i > 5$ (рис.12).

Окончательные вычисления показывают, что согласие распределений имеется (рис.13). Поскольку p-значение невелико, есть смысл применить другие критерии согласия, например, Шапиро-Уилка и Колмогорова-Смирнова (рис.14).

```
# Дополнительные критерии согласия
#Критерий Шапиро-Уилка для нормального закона распределения
shapiro.test(X)$p.value

## [1] 0.35121

#Критерий Колмогорова-Смирнова
ks.test((X-mu.)/sigma., "pnorm")

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  (X - mu.)/sigma.
## D = 0.053619, p-value = 0.6132
## alternative hypothesis: two-sided
```

Рис. 14. Проверка согласия по другим критериям.

Список литературы

1. Алексеева Н.П., Товстик Т.М. (2002) Практикум по математической статистике. СПб.: Изд-во Санкт-Петербургского университета..
2. Колемаев В.А., Калинина В.Н., Соловьев В.И. Малыхин В.И., Курочкин А.П. (2001) Теория вероятностей в примерах и задачах. Учебное пособие, ГУУ. – Москва, – 87 с.
3. Бородин А.Н. (2011) Элементарный курс теории вероятностей и математической статистики. Издательство «Лань» СПб, 256 с.
4. Крамер Г. (1975) Математические методы статистики. Изд-во "Мир Москва, - 648с.
5. Феллер В. (1984) Введение в теорию вероятностей и ее приложения. Том 1. Изд-во "Мир Москва, - 528с.