Оглавление

1. Описательная статистика и моделирование

1.1. 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 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.

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 $b \approx \frac{\omega}{k}$. 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 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 $\epsilon > 0$*

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

Действительно, если считать „успехом“ событие $x_i \le x$ с вероятностью $p = P\{x_i \le x\}$, 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 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

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

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

$$\begin{aligned} m_2 &= a_2 - \bar{x}^2 , \qquad (1)\\ m_3 &= a_3 - 3a_2\bar{x} + 2\bar{x}^3 ,\\ m_4 &= a_4 - 4a_2\bar{x} + 6a_2\bar{x}^2 - 3\bar{x}^4 . \end{aligned}$$

В частности

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

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

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

Утверждение 1. Второй выборочный центральный момент является смещенной оценкой дисперсии σ^2, то есть

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

Поэтому в качестве оценки дисперсии используют выражение

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

Для того чтобы убедиться в этом, введем $y_i = x_i - \mu$ с математическим ожиданием $\mathbf{E} y_i = 0$ и дисперсией $\mathbf{D} y_i = \sigma^2$, тогда $\bar{y} = \bar{x} - \mu$,

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

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

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

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

$$P\{\xi \leq x_P\} = P. \tag{3}$$

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 арифметическое.

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 $\frac{1}{h}$ и сосчитаем производные по h

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

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

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

При $h = 1$ из последнего уравнения получаем

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

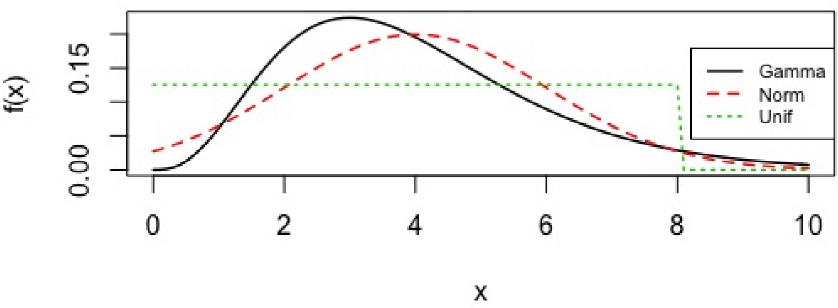

Рис. 1. Эксцесс нормального, равномерного и гамма распределений равны соответственно 0, $-\frac{6}{5}$ и $\frac{6}{\lambda}$, где λ параметр формы.

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,

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

1.3. Производящие функции моментов

Производящая функция определяется для целочисленной случайной величины ξ, принимающей значения $0, 1, 2, \ldots$ с вероятностями $p_0, p_1, p_2, \ldots$, $\sum_{j=1}^{\infty} p_j = 1$, как

$$h(\nu) = \sum_{j=0}^{\infty} p_j \nu^j \,. \tag{4}$$

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

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

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

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

Главное прикладное значение производящих функций заключается в представлении производящей функции суммы двух независимых величин в виде произведений частных производящих функций. Это свойство справедливо и для производящей функции моментов, которая определяется для случайной величины ξ (необязательно целочисленной) как

$$g(t) = \mathbb{E}e^{t\xi} = \begin{cases} \int\limits_{\Omega} e^{tx} f(x) dx & \text{в непрерывном случае,} \\ \sum_{j=1}^{\infty} p_j e^{tj} & \text{в дискретном случае}\,. \end{cases} \tag{5}$$

Значение k-й производной производящей функции моментов в нуле совпадает с k-м начальным моментом.

1.4. Практическое занятие по теме моделирования

На этом занятии предполагается рассмотреть основные законы распределения (биномиальное, Пуассона, равномерное, нормальное), вычислить математические ожидания, дисперсии, коэффициенты асимметрии и эксцесса. Промоделировать выборки, построить гистограммы и распределения модельных распределений. Сравнить характеристики теоретического и смоделированного распределения.

Биномиальное распределение

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**.

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

$$\xi : \begin{pmatrix} x_1 & x_2 & \ldots & x_N \\ p_1 & p_2 & \ldots & p_N \end{pmatrix}. \tag{6}$$

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**:

$$P\{\xi = k\} = C_n^k p^k (1-p)^{n-k},\ k = 0, 1, \ldots, n. \tag{7}$$

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

Дисперсия суммы двух случайных величин ξ, η равна

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

Если величины ξ, η независимы, то ковариация

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

и дисперсия суммы независимых величин равна сумме дисперсий. 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 СВ с распределением Бернулли $size = 1$.

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 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.

$$p_k = \frac{n(n-1)\ldots(n-k+1)}{k! n^k} (np)^k (1-p)^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda}, \qquad (8)$$

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

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

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

Случайная величина ξ, принимающая значения $k = 0, 1, 2, \dots$ с вероятностями (8), имеет **распределение Пуассона** $\mathcal{P}(\lambda)$. Так как $e^{\lambda} = \sum\limits_{k=0}^{\infty}\frac{\lambda^k}{k!}$, то $\sum\limits_{k=0}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda} = 1$. Математическое ожидание и дисперсия случайной величины ξ равны λ, так как

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

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

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

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

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.

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

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

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

Производящая функция моментов, 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 $runif(n, min = a, max = b)$.

Нормальное распределение

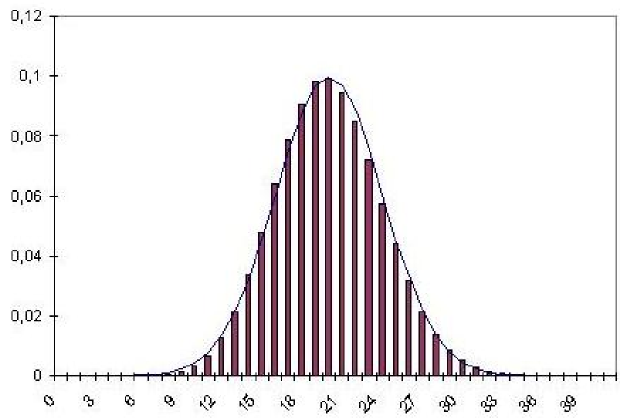

Рис. 2. Биномиальное распределение с параметрами $n = 100$ и $p = 0.2$ и нормальное распределение с параметрами $\mu = 20$, $\sigma^2 = 16$.

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

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{9}$$

При $\mu = 0$, $\sigma^2 = 1$ распределение $f(x|0,1)$ называется стандартным нормальным. Для доказательства основного свойства плотности нужно вычислить двойной интеграл

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

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

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

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

Отсюда $\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}$, и справедливо свойство плотности. Пусть $\xi_0 \sim \mathcal{N}(0,1)$. МО $\mathbb{E}\xi_0 = 0$ из-за симметричности функции относительно нуля:

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

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

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

$$= -\frac{2}{\sqrt{2\pi}} x e^{-\frac{x^2}{2}}\Big|_0^{\infty} + \frac{2}{\sqrt{2\pi}} \int\limits_{0}^{\infty} e^{-\frac{x^2}{2}} dx = 1\,.$$

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

Производящая функция моментов, 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 $rnorm(n, mean = \mu, sd = \sigma)$.

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

Таблица 2. Характеристики распределений

Покажем, что эксцесс равномерного распределения равен $-\frac{6}{5}$. Так как эксцесс не зависит от сдвига и масштаба, рассмотрим случай $\xi \sim U(0, 1)$.

$$\alpha_k = \int_0^1 x^k dx = \frac{1}{k+1}\,.$$

$$\begin{aligned} \mu_2 = \alpha_2 - \alpha_1^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}, \\ \mu_3 = 0, \\ \mu_4 = \alpha_4 - 4\alpha_3\alpha_1 + 6\alpha_2\alpha_1^2 - 3\alpha_1^4 = \frac{1}{5} - \frac{4}{4\cdot 2} + \frac{6}{3\cdot 4} - \frac{3}{16} = \frac{1}{80}, \\ \gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{12^2}{80} - 3 = -\frac{96}{80} = -\frac{6}{5}\,. \end{aligned}$$

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

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

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

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

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

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

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

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

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

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

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

Пример домашнего задания, расчитанного на несколько занятий. 1) Промоделировать 500 значений случайной величины, имеющей сложно-пуассоновское распределение с параметром $\lambda = 1$ пуассоновского числа нормально распределенных компонент с параметрами $\mu = 2, \sigma^2 = 0.24$. 2) Вычислить закон распределения, теоретические и выборочные характеристики распределения: среднее, дисперсию, асимметрию, эксцесс. 3) Оценить параметры распределения по выборке методом моментов и методом максимального правдоподобия. 4) Проверить согласие с моделью при помощи критерия хи-квадрат.