Вводная лекция по математической статистике

Н.П. Алексеева,

СПбГУ, математико-механический факультет
5665 «Математическое обеспечение и администрирование информационных систем»

2021 г.

Главная задача математической статистики и понятие выборки

Главная задача математической статистики

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

Выборка

Конечная совокупность наблюдений $x_1, \ldots, x_n$ над случайной величиной ξ называется выборкой.

Очевидно, что компоненты x_j, $j = 1, 2, \ldots, n$, — независимые в совокупности и распределенные по тому же закону, что и случайная величина ξ.

Типы признаков и примеры выборок

Случайным величинам, как правило, соответствуют некоторые признаки, которые подразделяются на

- количественные – рост школьников в см: $(132, 134, 134, 136, 143, \ldots)$ — когда способ измерения в некоторой шкале;
- порядковые — успеваемость: (высокая, высокая, средняя, низкая, средняя,...) — измерить нельзя, но можно упорядочить;
- качественные — цвет глаз: (карий, карий, зеленый, серый, карий, голубой, ...) — когда градации можно менять местами.

p-мерную выборку представляют в виде следующей таблицы.

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

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

Эмпирическое распределение

Закон распределения случайной величины ξ называется законом распределения генеральной совокупности. Реализация выборочного вектора называется выборкой $(x_1, \ldots, x_n)$ объема n.

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 $1/n$.

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

$$\text{успеваемость} : \begin{pmatrix} \text{высокая} & \text{средняя} & \text{низкая} \\ 5 & 12 & 13 \end{pmatrix}$$

В случае метрической или порядковой выборки соответствующая функция распределения называется выборочной или эмпирической функцией распределения

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

где $\mu_n(x) = \sharp\{x_j : x_j \le x\}$ равно количеству элементов выборки, не больших x.

Теорема (Гливенко)

Для любых $x \in (-\infty; +\infty)$ и $\epsilon > 0$

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

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.

Доказательство теоремы Гливенко

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

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

Теорема Бернулли доказывается на основе неравенства Бьеномэ-Чебышева:

$$P\{|\xi - \mathbb{E}\xi| < \varepsilon\} \geq 1 - \frac{\mathbb{D}\xi}{\varepsilon^2} \iff P\left\{\left|\frac{\mu_n(x)}{n} - p\right| < \epsilon\right\} \geq 1 - \frac{pq}{n\varepsilon^2},$$

$$\text{так как } \mathbb{D}\left(\frac{\mu_n(x)}{n}\right) = \frac{npq}{n^2} = \frac{pq}{n}.$$

Неравенство Бьеномэ-Чебышева

$$\mathrm{P}\{|\xi - \mathbb{E}\xi| < \varepsilon\} \geq 1 - \frac{\mathbb{D}\xi}{\varepsilon^2}$$

Пусть S — множество точек, удовлетворяющих $\mathrm{g}(\xi) \geq \mathrm{K}$, то

$$\mathbb{E}\mathrm{g}(\xi) = \int\limits_{-\infty}^{+\infty} \mathrm{g(x)dF} \geq \mathrm{K} \int_{\mathrm{S}} \mathrm{dF} = \mathrm{KP(S)} \iff \mathrm{P}\{\mathrm{g}(\xi) \geq \mathrm{K}\} \leq \frac{\mathbb{E}\mathrm{g}(\xi)}{\mathrm{K}}$$

Пусть $\mathrm{g}(\xi) = |\xi - \mathbb{E}\xi|$

$$\mathrm{P}\{|\xi - \mathbb{E}\xi| \geq \varepsilon\} = \mathrm{P}\{(\xi - \mathbb{E}\xi)^2 \geq \varepsilon^2\} \leq \frac{\mathbb{D}\xi}{\varepsilon^2},$$

$$\mathrm{P}\{|\xi - \mathbb{E}\xi| < \varepsilon\} = 1 - \mathrm{P}\{|\xi - \mathbb{E}\xi| \geq \varepsilon\} \geq 1 - \frac{\mathbb{D}\xi}{\varepsilon^2}\,.$$

Характеристики выборки

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.

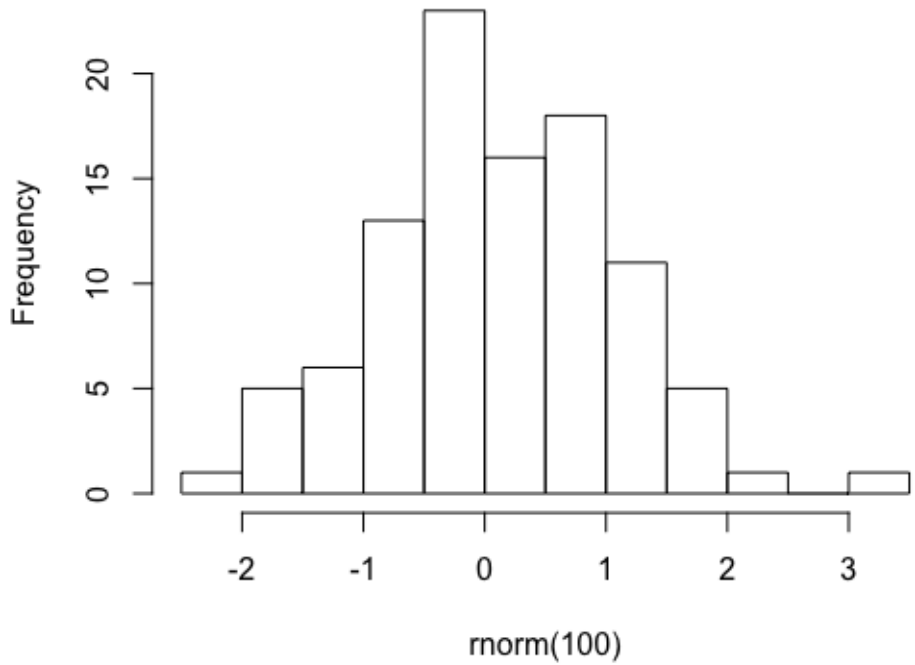
Группированный статистический ряд

При большом объеме выборки ее элементы объединяются в группы (разряды, карманы и т.п.), представляя результаты в виде группированного статистического ряда.

- k — число непересекающихся интервалов;
- ν_i — частоты, равные количеству элементов выборки, попадающих в этот интервал;
- z_i — середины интервалов группировки,
- $b \approx \frac{\omega}{k}$ — длина интервала группировки;
- $\frac{\nu_i}{n}$ — относительные частоты.

Гистограмма

Гистограммой частот группированной выборки будем называть кусочно-постоянную функцию, принимающую на интервалах группировки значения n_i. Гистограмма относительных частот определяется аналогично высотой $\frac{n_i}{nb}$ и единичной площадью под графиком. Функция hist() в R.



Полигон частот

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

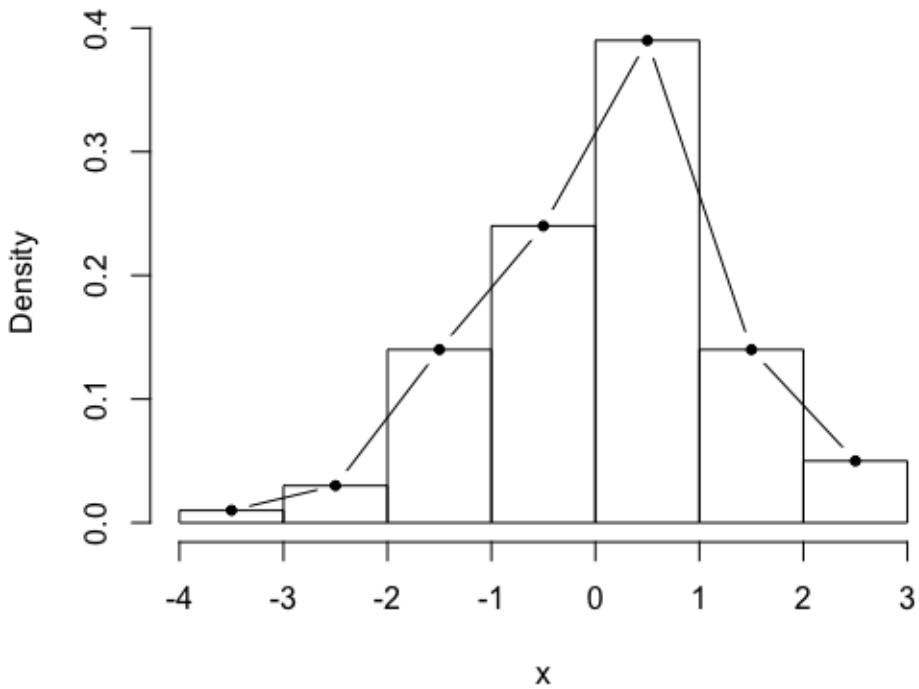


Figure: $y <- \text{hist}(x, \text{plot} = \text{FALSE}); \text{plot}(y, \text{freq} = \text{FALSE}, \text{main} = \text{NULL}); \text{lines}(y\$\text{mids}, y\$\text{density}, \text{type} = 'b', \text{pch} = 20)$

Характеристики выборочного распределения

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

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

В частности

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

$$= a_2 - 2\bar{x}\bar{x} + \bar{x}^2 = a_2 - \bar{x}^2\,,$$

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3 =$$

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

$$= a_3 - 3a_2\bar{x} + 3\bar{x}^2\bar{x} - \bar{x}^3 = a_3 - 3a_2\bar{x} + 2\bar{x}^3\,.$$

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i^4 - 4x_i^3\bar{x} + 6x_i^2\bar{x}^2 - 4x_i\bar{x}^3 + \bar{x}^4) =$$

$$= a_4 - 4a_3\bar{x} + 6a_2\bar{x}^2 - 4\bar{x}^3\bar{x} + \bar{x}^4 = a_4 - 4a_3\bar{x} + 6a_2\bar{x}^2 - 3\bar{x}^4$$

Свойства оценок: несмещенность

Оценка параметра $\hat{\theta}_n$ называется несмещенной, если ее математическое ожидание совпадает с истинным значением параметра, т.е. $E\hat{\theta}_n = \theta$. Если $E\hat{\theta}_n = \theta + b(\theta)$, то $b(\theta)$ называется смещением.

Пример несмещенной оценки математического ожидания μ генеральной совокупности с дисперсией σ^2 в виде $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Используем независимость сл.вел. x_j и то, что $\mathbb{E}x_j = \mu$ для любого $j = 1, \ldots, n$.

$$\mathbb{E}\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}x_i = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu,$$

$$\mathbb{D}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{D}x_i = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}.$$

Пример смещенной оценки

Второй выборочный центральный момент $m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ является смещенной оценкой дисперсии σ^2, то есть

$$Em_2 = \frac{\sigma^2(n-1)}{n}. \tag{2}$$

Введем $y_i = x_i - \mu$ с $Ey_i = 0$ и дисперсией $Dy_i = \sigma^2$, тогда $\bar{y} = \bar{x} - \mu$,

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

$$Em_2 = \frac{1}{n}\sum_{i=1}^{n} Ey_i^2 - E\bar{y}^2 = \sigma^2 - \frac{1}{n^2}\sum_{i,j} Ey_iy_j = \sigma^2 - \frac{1}{n^2}\sum_{i} Ey_i^2 = \sigma^2 - \frac{\sigma^2}{n},$$

так как $Ey_i^2 = Dy_i = \sigma^2$, а при $i \neq j$ из-за независимости элементов выборки $Ey_iy_j = Ey_iEy_j = 0$.

Несмещенная оценка дисперсии

В качестве оценки дисперсии используют выражение

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

$$\mathbb{E}S^2 = \frac{n}{n-1} \mathbb{E}m_2 = \frac{n}{n-1} \left(\sigma^2 - \frac{\sigma^2}{n} \right) = \sigma^2 .$$

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

Мода

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.

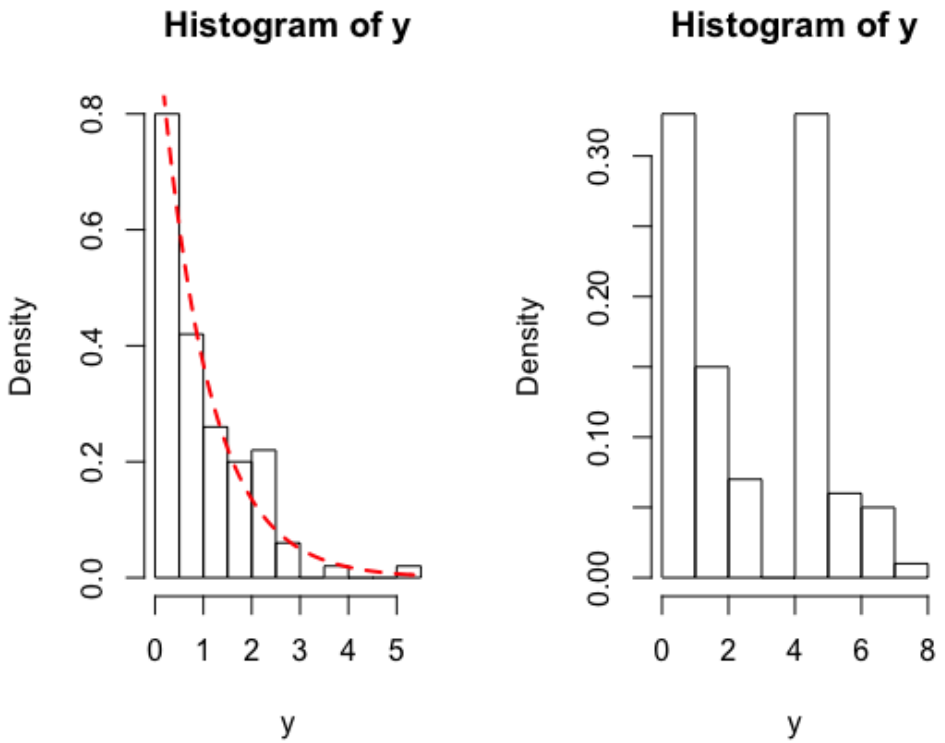
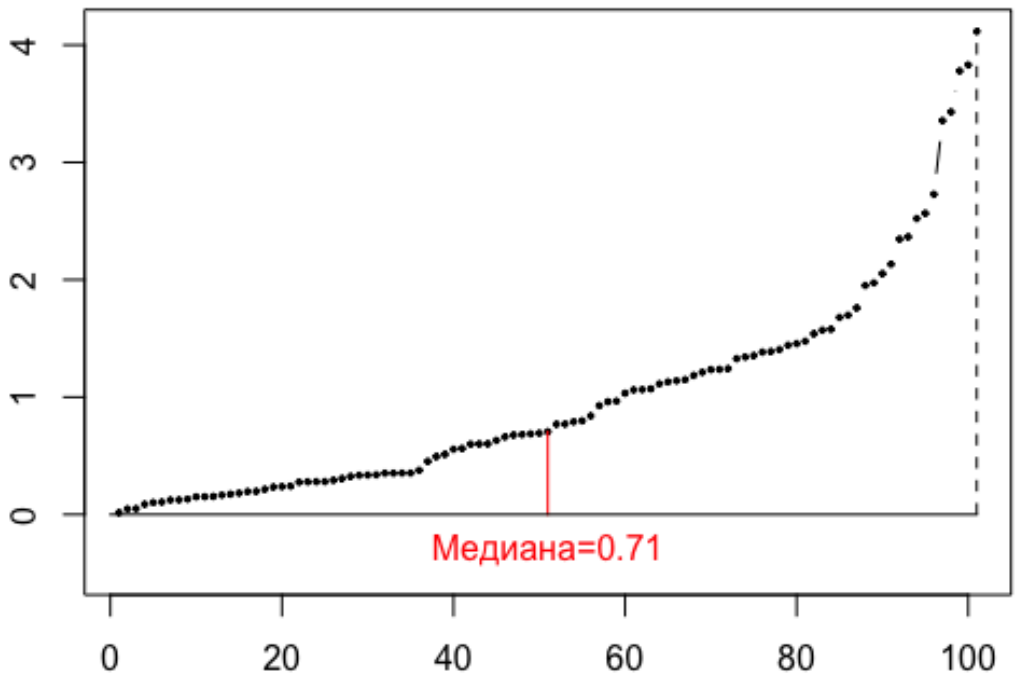


Figure: Унимодельное и бимодальное распределение.

Медиана

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 примере $1, 1, 1, 1, 2, 2, 2, 3, 3, 4$ $med = 2$.

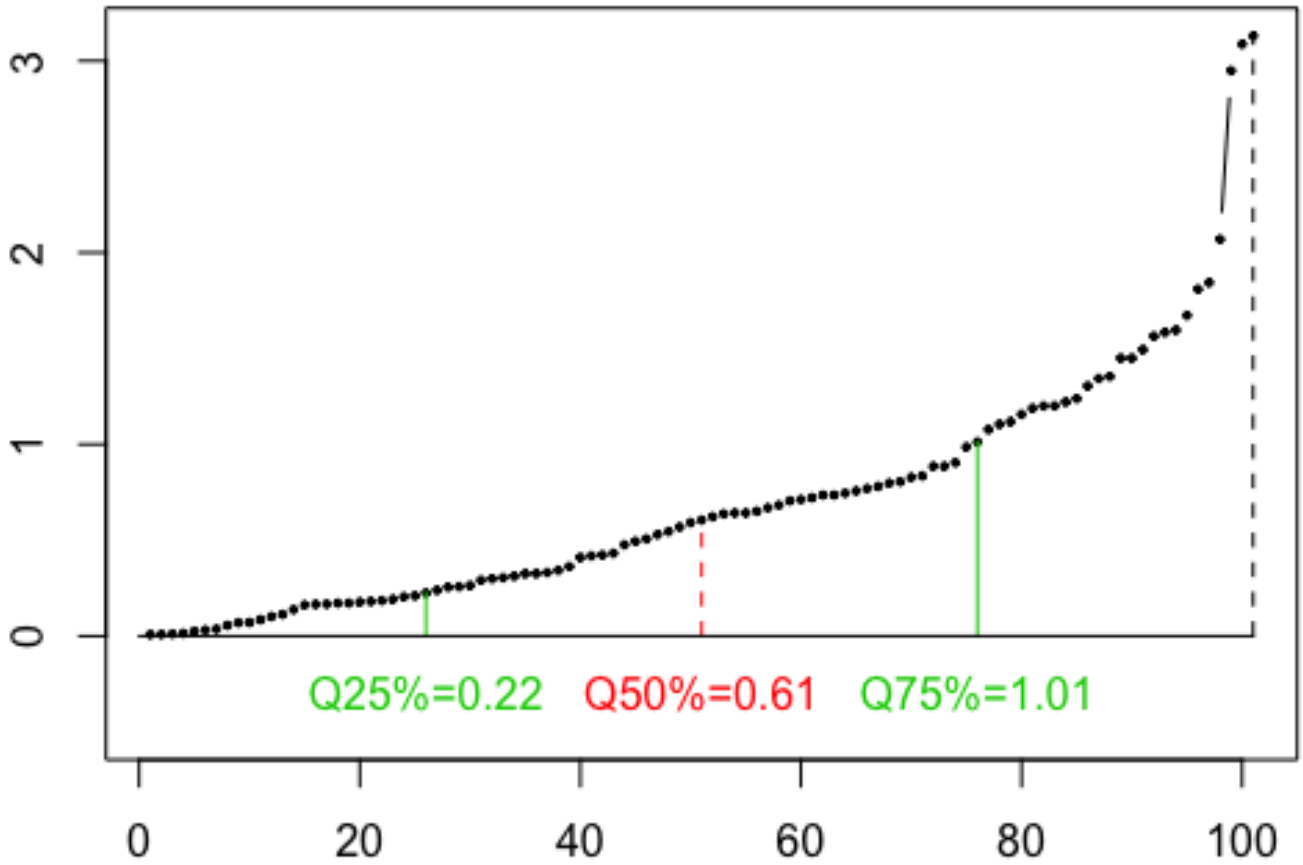


P-квантиль

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 арифметрическое.
Для описания переменных используется сочетание: минимум, 25%квантиль, медиана, 75%квантиль, максимум.


0
1
2
3
0
20
40
60
80
100
Q25%=0.22
Q50%=0.61
Q75%=1.01

Коэффициент асимметрии $\gamma_1 = \mu_3/\mu_2^{1.5}$

В случае симметричного закона распределения все нечетные моменты равны нулю.

Для симметричной плотности распределения $f(E\xi - x) = f(E\xi + x)$.

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

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

В первом интеграле замена $x - E\xi = y$, а во втором $x - E\xi = -y$.

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

Выборочный коэффициент асимметрии $g_1 = m_3/m_2^{1.5}$

$$g_1(aX + b) = \frac{m_3(aX + b)}{m_2(aX + b)^{1.5}} = \frac{a^3 m_3(X)}{(a^2 m_2(X))^{1.5}} = g_1(X)$$

Эксцесс $\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$ как метрика отклонения от нормального закона распределения

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

В нормальной плотности заменим σ^2 на $\frac{1}{\mathrm{h}}$ и продифференцируем по h

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

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

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

при $\mathrm{h}=1$ $\displaystyle\int\limits_{-\infty}^{+\infty} \mathrm{e}^{\frac{-\mathrm{t}^2}{2}}\left(\frac{\mathrm{t}^4}{4}\right)\mathrm{dt} = \sqrt{2\pi}\left(\frac{3}{4}\right), \implies \mu_4 = 3.$

Эксцесс нормального распределения

Если $\xi \sim \mathcal{N}(0,1)$, то 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,

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

Выборочный эксцесс вычисляется как $g_2 = \frac{m_4}{m_2^2} - 3$.

Эксцессы разных распределений

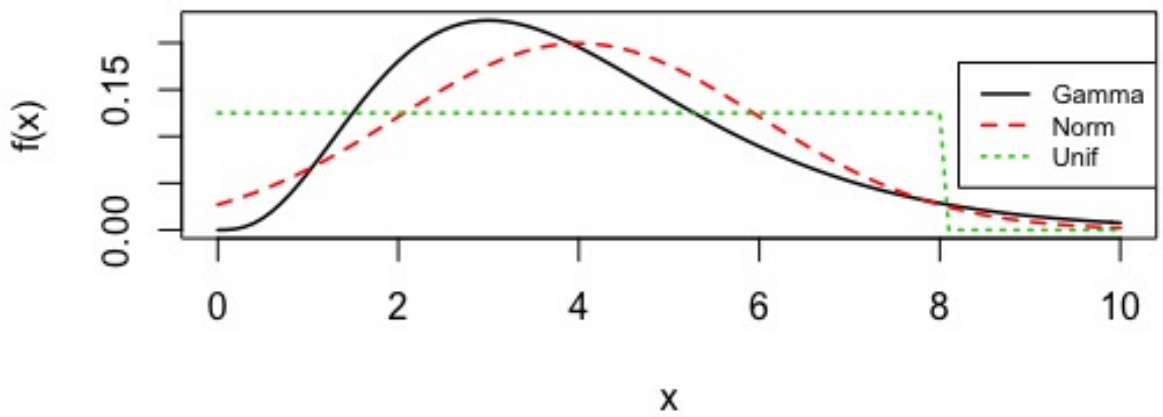


Figure: Эксцессы нормального, равномерного и гамма распределений равны соответственно 0, $-\frac{6}{5}$ и $\frac{6}{\lambda}$, где λ параметр формы.

Примеры вычисления асимметрии и эксцесса

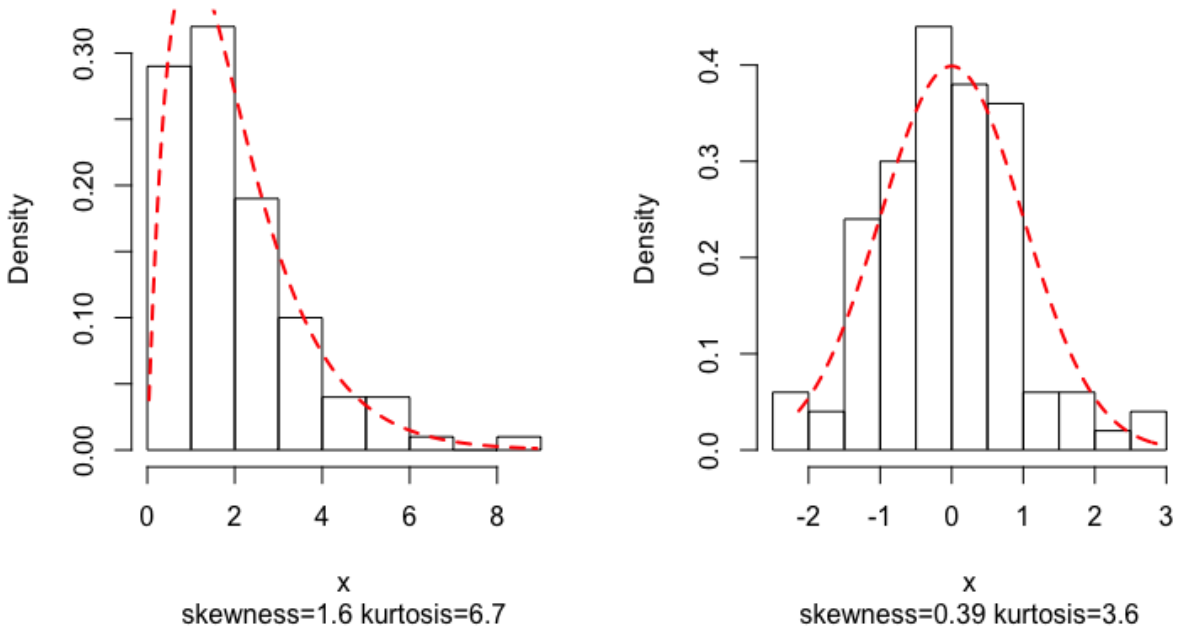


Figure: Сравнение гистограмм гамма (параметр формы 2 и параметр масштаба 1) и нормального распределений.

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.