ОСНОВЫ ХИМИЧЕСКОЙ МЕТРОЛОГИИ
Лекция 2 Эмпирические и теоретические распределения

лектор: Образовский Е. Г.

17 июля 2016 г.

Эмпирические распределения

Опыт показывает, что результаты количественного химического анализа (КХА) являются случайными величинами. Случайные погрешности, характеризующие прецизионность анализа, имеют вероятностную природу и описываются с помощью методов математической статистики.

В процессе измерения случайная величина принимает какое-либо одно значение из допустимого набора значений. Для полной характеристики необходимо знать вероятность появления тех или иных значений.

Распределением (вероятностей) называется функция, определяющая вероятность того, что случайная величина примет какое-либо значение или будет принадлежать заданному множеству значений.

Большое число экспериментальных данных удобно представлять в виде гистограммы. Наиболее наглядная картина распределения результатов анализа получится, если разбить на интервалы область значений (число интервалов k выбирается $k \sim \sqrt{n}$, где n – полное число результатов анализа; другой вариант $k = 1 + 3,32 \cdot \lg n$) и построить гистограмму – зависимость числа результатов анализа в данном интервале от номера интервала. На рис. 3 представлены в виде гистограмм данные межлабораторных сравнительных испытаний по определению иона аммония в природной воде, полученных в 79 лабораториях. Увеличение числа интервалов приводит к сильным флуктуациям числа результатов в отдельных интервалах и поэтому нецелесообразно.

Эмпирические распределения

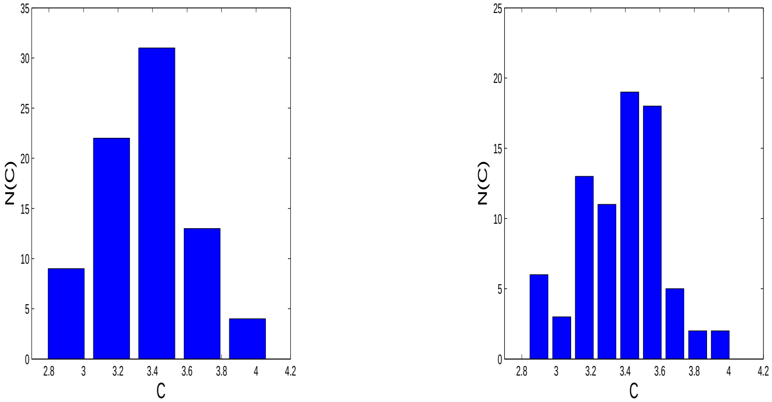


Рис.: Представление эмпирических данных в виде гистограмм для числа интервалов $k = 5, 9$

Эмпирические распределения

Наиболее важные характеристики эмпирического распределения – среднее арифметическое (центр рассеяния результатов)

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

и стандартное отклонение (степень рассеяния результатов, т. е. ширина распределения)

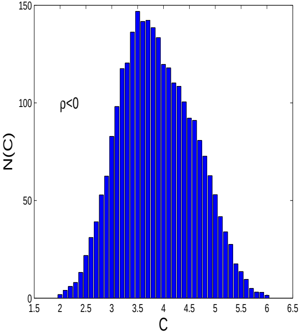
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Не все эмпирические распределения схожи с симметричным нормальным распределением. Количественно асимметрию распределения характеризуют параметром

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n \cdot s^3}.$$

Если $\rho > 0$ – левосторонняя асимметрия (меньшее число результатов анализа, для которых $x_i < \bar{x}$), если $\rho < 0$, то правосторонняя асимметрия (большее число результатов, для которых $x_i < \bar{x}$).

Эмпирические распределения



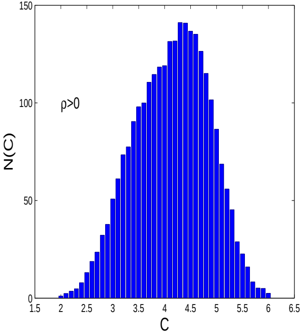


Рис.: Асимметричные распределения

Симметричные распределения также могут отклоняться от нормального закона распределения. Нарушение случайности выборки (предвзятость, стремление получить "нужное значение") приводит к островершинным распределениям. Если объединить результаты анализа, полученные в существенно разных условиях, то у распределения получаются пологие (плоские) максимумы. Количественно эти эффекты характеризуются эксцессом ε

$$\varepsilon = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n \cdot s^4} - 3.$$

Островершинному распределению соответствует значение $\varepsilon > 0$, пологому $\varepsilon < 0$, для нормального распределения $\varepsilon = 0$.

Эмпирические распределения

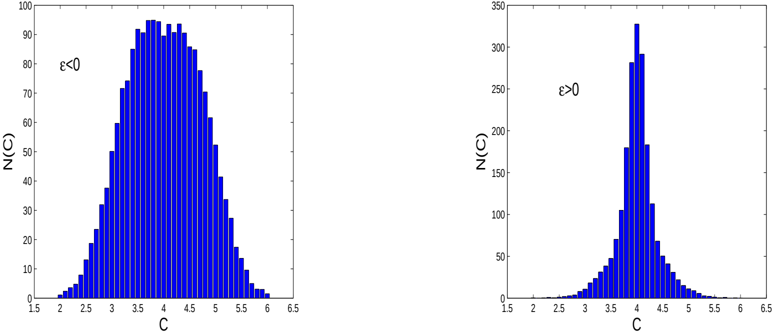


Рис.: Пологое и островершинное распределения

Нормальное распределение

На основании опытных данных, как правило, принимают, что распределение совокупности результатов КХА соответствует так называемому нормальному распределению – распределению Лапласа–Гаусса.

Рис.: К. Ф. Гаусс

Нормальное распределение

Теоретическое обоснование: суммарная погрешность анализа складывается из большого числа независимых вкладов, так что, согласно центральной предельной теореме, каков бы ни был закон распределения погрешностей на каждом этапе, закон распределения суммарной погрешности стремится к нормальному при достаточно большом числе составляющих вкладов.

Рис.: П. С. Лаплас

Нормальное распределение

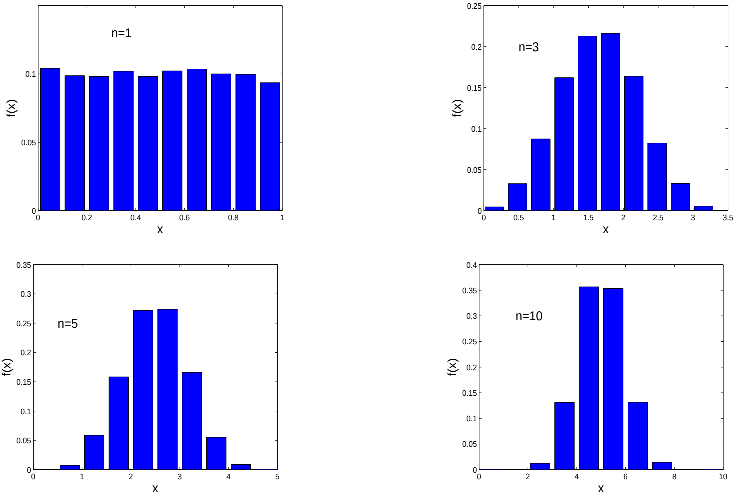


Рис.: Распределения суммы $n = 1, 3, 5, 10$ равномерно распределенных случайных чисел при увеличении n стремится к нормальному распределению

Нормальное распределение

Математически функция нормального распределения описывается формулой Гаусса

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

где $f(x, \mu, \sigma)$ есть плотность распределения (рис. 7). Вероятность $P(x_1 \leqslant x \leqslant x_2)$ найти значение результата анализа в интервале (x_1, x_2) есть

$$\mathbf{P}(\mathbf{x_1} \leqslant \mathbf{x} \leqslant \mathbf{x_2}) = \int_{\mathbf{x_1}}^{\mathbf{x_2}} \mathbf{f}(\mathbf{x}, \mu, \sigma)\mathbf{dx}.$$

Нормальное распределение

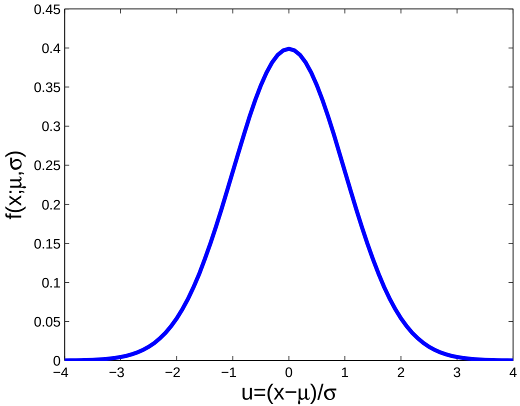


Рис.: Нормальное (гауссово) распределение

Нормальное распределение

Это распределение одномодальное (имеет один максимум) и характеризуется двумя параметрами: μ – средним значением переменной x

$$\mathbf{\bar{x}} = \int_{-\infty}^{+\infty} \mathbf{x} \cdot \mathbf{f}(\mathbf{x}, \mu, \sigma)\mathbf{dx} = \mu,$$

совпадающим с наиболее вероятным значением плотности распределения, и стандартным отклонением σ, характеризующим ширину распределения

$$\overline{(\mathbf{x} - \mathbf{\bar{x}})^2} = \int_{-\infty}^{+\infty} (\mathbf{x} - \mathbf{\bar{x}})^2 \cdot \mathbf{f}(\mathbf{x}, \mu, \sigma)\mathbf{dx} = \sigma^2.$$

Нормальное распределение

Вероятность $P = 1 - \alpha$ найти значение x внутри интервала $(\mu - \Delta, \mu + \Delta)$ определяется интегралом

$$\mathbf{P} = \mathbf{1} - \alpha = \int_{\mu-\boldsymbol{\Delta}}^{\mu+\boldsymbol{\Delta}} \mathbf{f}(\mathbf{x}, \mu, \sigma)\mathbf{dx}.$$

При увеличении ширины интервала эта вероятность очень быстро приближается к $100\ \%$. Так, для $\Delta = 1,96 \cdot \sigma$ вероятность равна $95\ \%$; для $\Delta = 3,29 \cdot \sigma$ вероятность равна $99,9\ \%$

Законы сложения ошибок Для получения конечного результата анализа часто приходится производить математические операции над случайными величинами.

а) *Сложение и вычитание* Найдем плотность распределения $f(z)$ суммы $z = x + y$ двух нормально распределенных величин с параметрами распределения соответственно μ_1, σ_1 и сμ_2, σ_2:

$$\mathbf{f(z)} = \int_{-\infty}^{+\infty} \mathbf{f_1(x}, \mu_1, \sigma_1\mathbf{)f_2(z - x}, \mu_2, \sigma_2\mathbf{)dx} = \mathbf{N}\mathbf{e}^{-(\mathbf{z}-\mu_1-\mu_2)^2/2(\sigma_1^2+\sigma_2^2)},$$

где нормировочный множитель $N = (2\pi(\sigma_1^2 + \sigma_2^2))^{-1/2}$.

Нормальное распределение

Моделирование распределения для суммы двух нормальных распределений.

Нормальное распределение

Моделирование распределения для суммы двух нормальных распределений.

Нормальное распределение

Иначе говоря, для суммы получается среднее значение $\bar{z} = \bar{x} + \bar{y} = \mu_1 + \mu_2$ и стандартное отклонение $\sigma_z = \sqrt{\sigma_1^2 + \sigma_2^2}$.

Следствие. Для двух результатов измерения x_1 x_2 одной и той же величины стандартное отклонение среднеарифметического есть $\sigma_{\bar{x}} = \sigma_x/\sqrt{2}$; для среднеарифметического n-результатов – $\sigma_x/\sqrt{n}$.

Нормальное распределение

Моделирование распределения для полусуммы двух нормальных распределений.

Нормальное распределение

Моделирование распределения для полусуммы двух нормальных распределений.

Нормальное распределение

Для разности $z = x - y$ получается $\bar{z} = \mu_1 - \mu_2$, $\sigma_z = \sqrt{\sigma_1^2 + \sigma_2^2}$.

Нормальное распределение

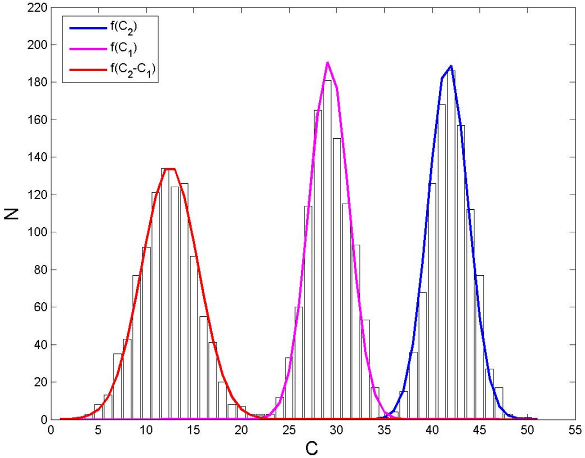
Моделирование распределения для разности двух нормальных распределений.

Нормальное распределение

Моделирование распределения для разности двух нормальных распределений.

Нормальное распределение

В качестве примера приведем распределение разности двух нормально распределенных величин



б) *Умножение и деление* В данном случае результат имеет нормальное распределение только при условии $\sigma_1/\mu_1 \ll 1$, $\sigma_2/\mu_2 \ll 1$. Тогда

$$\mathbf{z} = \mathbf{x} \cdot \mathbf{y}, \quad \bar{\mathbf{z}} = \bar{\mathbf{x}} \cdot \bar{\mathbf{y}} = \mu_1 \cdot \mu_2, \quad \frac{\sigma_z}{\mu_1 \mu_2} = \sqrt{\left(\frac{\sigma_1}{\mu_1}\right)^2 + \left(\frac{\sigma_2}{\mu_2}\right)^2},$$

$$\mathbf{z} = \frac{\mathbf{x}}{\mathbf{y}}, \quad \bar{\mathbf{z}} = \frac{\bar{\mathbf{x}}}{\bar{\mathbf{y}}} = \frac{\mu_1}{\mu_2}, \quad \frac{\sigma_z}{\mu_1/\mu_2} = \sqrt{\left(\frac{\sigma_1}{\mu_1}\right)^2 + \left(\frac{\sigma_2}{\mu_2}\right)^2}.$$

Нормальное распределение

В общем случае функциональной зависимости

$$\phi = \phi(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n})$$

от n независимых нормально распределенных величин при условии $\sigma_i/\mu_i \ll 1$, $i = 1, ..., n$ погрешность определяется выражением

$$\delta\phi \approx \frac{\partial\phi}{\partial\mathbf{x_1}} \cdot \delta\mathbf{x_1} + ... + \frac{\partial\phi}{\partial\mathbf{x_n}} \cdot \delta\mathbf{x_n}.$$

Нормальное распределение

При условии независимости переменных x_i, что математически выражается как

$$\overline{\delta x_i \delta x_j} = \iint (x_i - \mu_i)(x_j - \mu_j) f_i(x_i, \mu_i, \sigma_i) f_j(x_j, \mu_j, \sigma_j) dx_i dx_j = 0,$$

стандартное отклонение для функции ϕ равно [1]

$$\sigma_\phi = \sqrt{\overline{\delta\phi^2}} = \sqrt{\left(\frac{\partial\phi}{\partial x_1}\right)^2 \cdot \sigma_1^2 + \ldots + \left(\frac{\partial\phi}{\partial x_n}\right)^2 \cdot \sigma_n^2}.$$

[1] Приведенные законы сложения ошибок верны не только для нормального распределения.

Пример: фотометрия Обозначим I_0 – интенсивность света при отсутствии пробы, I – интенсивность при наличии пробы. Концентрацию определяемого компонента вычисляют по значению оптической плотности D по формуле

$$\mathbf{D} = \lg\left(\frac{\mathbf{I_0}}{\mathbf{I}}\right) = \phi(\mathbf{I_0}, \mathbf{I}).$$

Считая стандартные отклонения интенсивностей приближенно равными $\sigma_{I_0} \approx \sigma_I = \sigma$ и получаем для относительного стандартного отклонения оптической плотности

$$\frac{\sigma_\mathbf{D}}{\mathbf{D}} = \frac{\sigma}{\mathbf{I_0} \cdot \lg(\mathbf{I_0}/\mathbf{I})}\sqrt{\mathbf{1} + \left(\frac{\mathbf{I_0}}{\mathbf{I}}\right)^{\mathbf{2}}}.$$

Эта величина минимальна при $I/I_0 \approx 0,33$.

Нормальное распределение

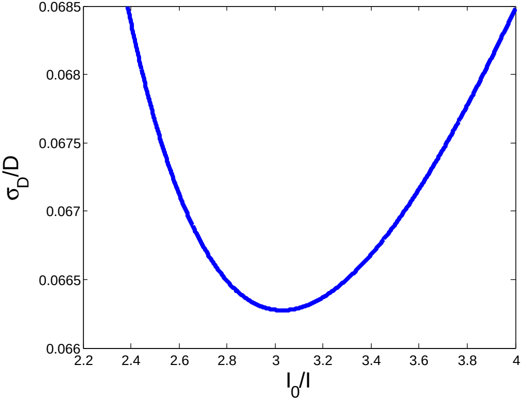


Рис.: Зависимость относительного стандартного отклонения оптической плотности σ_D/D от отношения интенсивностей I_0/I

Распределение Пуассона

В некоторых методах анализа (например, в рентгенофлуоресцентном) аналитический сигнал является дискретной величиной (число импульсов в полупроводниковом или сцинтилляционном детекторе и т. д.). При заданной средней скорости счета I среднее ожидаемое число импульсов в детекторе за время t равно $\mu = I \cdot t$. Из-за случайного характера процессов испускания и регистрации излучения экспериментально получаемые значения числа импульсов имеют разброс. Вероятность получить число импульсов n описывается функцией распределения Пуассона

Рис.: С. Пуассон

$$\mathbf{f}(\mathbf{n}, \mu) = \frac{\mu^{\mathbf{n}}}{\mathbf{n}!} \exp(-\mu).$$

Распределение Пуассона

Это одномодальное (с одним максимумом) однопараметрическое распределение. Среднее значение $\bar{n} = \mu$, стандартное отклонение $\sigma = \sqrt{\mu}$.
При малых значениях μ наблюдается значительная асимметрия распределения. При $\mu \geq 10$—15 распределение Пуассона переходит в нормальное распределение

$$\mathbf{f(n)} = \frac{\mathbf{1}}{\sqrt{\mathbf{2}\pi\mu}} \exp\left[-\frac{(\mathbf{n}-\mu)^{\mathbf{2}}}{\mathbf{2}\mu}\right].$$

Как и для нормального распределения, сумма и разность величин, подчиняющихся распределению Пуассона, подчиняются также распределению Пуассона.

Распределение Пуассона

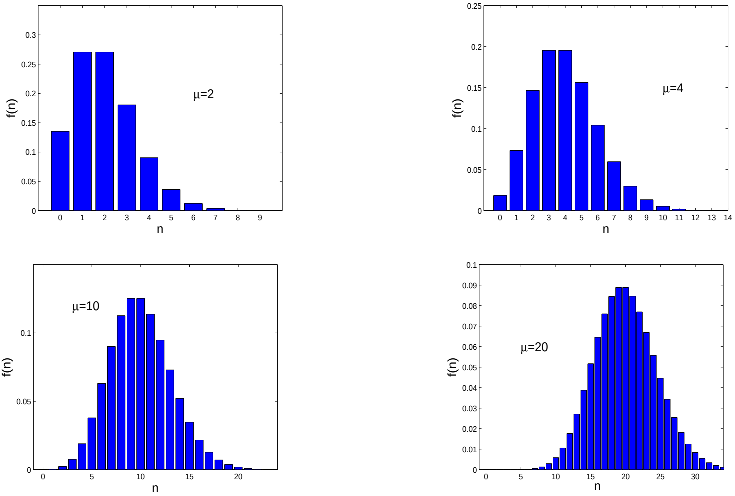


Рис.: Распределение Пуассона для $\mu = 2, 4, 10, 20$.

Пример. Если, как это обычно имеет место, кроме полезного сигнала имеется фоновый, то число импульсов полезного сигнала N_p находится как разность $N_p = N - N_f$, где N – суммарное число импульсов в области пика (сигнал + фон), N_f – число импульсов фона. Тогда стандартное отклонение для сигнала σ_p находится по закону сложения ошибок как

$$\sigma_\mathbf{p} = \sqrt{\sigma_\mathbf{N}^2 + \sigma_\mathbf{f}^2} \approx \sqrt{\mathbf{N} + \mathbf{N_f}} \approx \sqrt{\mathbf{N_p} + 2\mathbf{N_f}}.$$

Распределение Пуассона

Моделирование $\gamma-$спектра

Распределение Пуассона

Если ожидаемое среднее значение μ мало, то границы доверительного интервала при оценке μ по измеренному числу импульсов n становятся существенно асимметричными. Так, для наблюдаемого числа отсчетов $n = 9$ границы доверительного интервала для вероятности $P = 0,95$ согласно распределению Пуассона равны $(4,75; 16,77)$, тогда как нормальное распределение дает $(3,12; 14,88)$.