Оценка погрешностей будущих измерений по имеющимся данным

Поляченко Юрий

10 апреля 2020 г.

1 Постановка задачи

Цель – предсказать погрешность ε выдаваемых нашей программой значений искомого параметра x.

Область изменения $x \in [0;1]$ разбита на интервалы $[a_j;b_j]$, для каждого из которых есть N_j экспериментов. Считается, что изкомая погрешность ε может меняться от интервала к интервалу, но постоянная внутри интервала. Фиксируем j и работаем в выбранном интервале, поэтому далее его индекс опущен.

Есть N экспериментов, про которые известно, что в каждом из них истинное значения x попало в интервал $[a;b]$. На каждый i-ый из этих экспериментов у нас есть результат работы нашей программы x_i. Предполагается, что истинное значение y_i распределено по гауссу со средним x_i и дисперсией ε. Ищем зависимость

$$\varepsilon\left(a, b, \{x_i\}_{i=1}^{N}, p_0\right) \tag{1}$$

такую, что

$$\mathcal{P}(\forall x \in [a;b]\ \ |x-y| < \varepsilon) = p_0 \tag{2}$$

Из сторонних соображений считается известным минимально возможная погрешность ε_{min}.

2 Предлагаемое решение

2.1 Приближение

Задав ε, можно посчитать вероятность реализации ситуации, описанной в постановке – попадание всех истинных значений y_i параметра, распределенных согласно результатам работы нашей программы по гауссу каждый около своего x_i, в интервал $[a;b]$. Далее предположение - эта вероятность равна нашей целевой вероятность p_0. Не очевидно, почему это должно выполняться точно (скорее всего это не выполняется), но для оценки предложено использовать такую модель.

2.2 Расчет

Вероятность попадание i-ой истинной точки в интервал

$$p_i(\varepsilon) = \int\limits_a^b g\left(x, x_i, \varepsilon\right) dx = \int\limits_{(a-x_i)/\varepsilon}^{(b-x_i)/\varepsilon} g(x, 0, 1) dx, \tag{3}$$

где

$$g(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{4}$$

– гауссово распределение.
Введем

$$\mathsf{erf}(x) = \int\limits_{-\infty}^{x} g(t, 0, 1) dt. \tag{5}$$

Попадание каждого истинного значения в интервал - независимое событие, поэтому вероятность реализации нашего случая

$$p(\varepsilon) = \prod_{i=1}^{N} p_i(\varepsilon) = \prod_{i=1}^{N} \left[\mathsf{erf}\left(\frac{b-x_i}{\varepsilon}\right) - \mathsf{erf}\left(\frac{a-x_i}{\varepsilon}\right)\right] \tag{6}$$

Для нахождения желаемого ε_0 решаем уравнение $p(\varepsilon_0) = p_0$.
Очевидно, что

$$\left.\begin{aligned} \forall \varepsilon > 0 \quad p'(\varepsilon) < 0 \\ \mathrm{ran}[p(\varepsilon)] = (0; 1) \end{aligned}\right\} \Rightarrow \quad \forall p_0 \in (0; 1) \ \ \exists!\, \varepsilon > 0 : p(\varepsilon) = p_0, \tag{7}$$

поэтому $\forall p_0 \in (0; 1)$ уравнение хорошо решается численно.

2.3 Пример

Для примера можно взять случайный набор из 10 точек в интервале $[0.25 \cdot 1.05;\ \ 0.5 \cdot 0.95]$ и считать, что их истинные значения принаждежат $[0.25; 0.5]$.

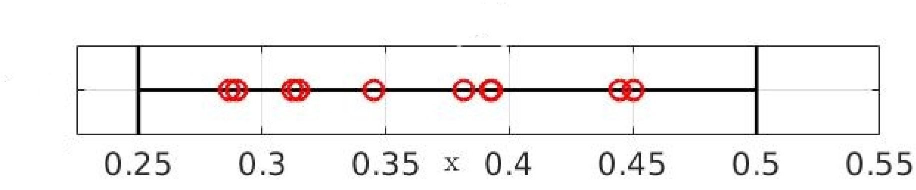

Рис. 1: Расположение 10 пробных точек в интервале $[0.25; 0.5]$.

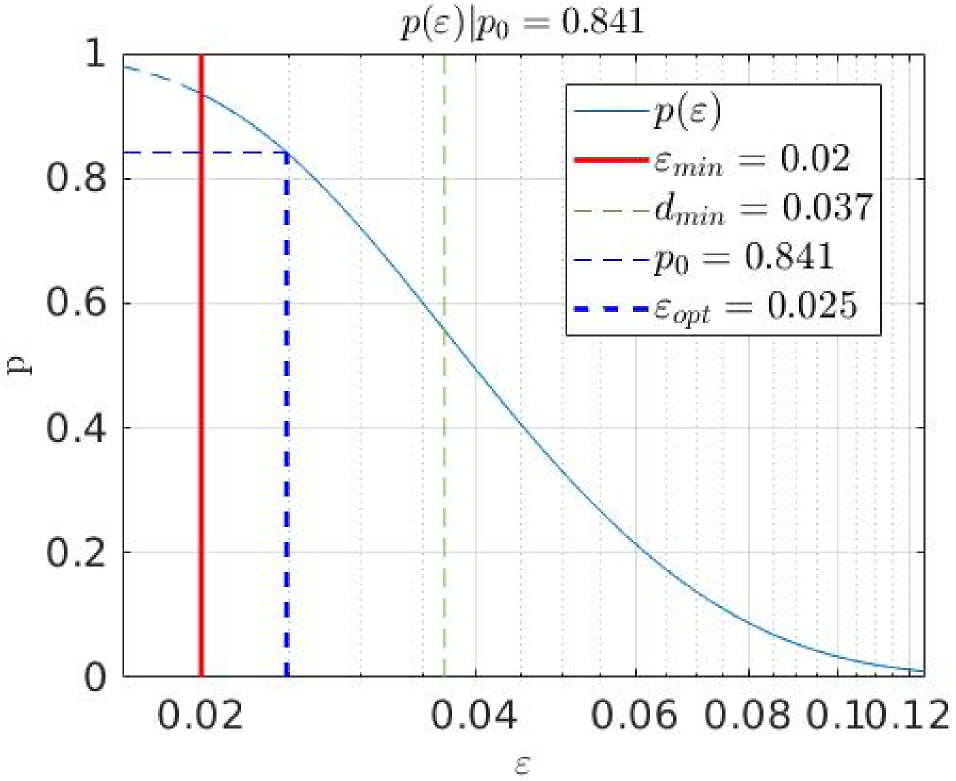

Рис. 2: Зависимость $p(\varepsilon)$. Красная линия – принятая минимально возможная погрешность $\varepsilon_{min} = 0.02$, Синяя вертикаль – найденная оценка, синяя горизонталь –наш выбор $p_0 = (1 - \mathcal{P}_{gauss}(1\sigma))/2$, зеленый пунктир - минимальное расстояние точек до границы. Видно, что наличие множества точке позволяет улучшить оценку с очевидного значения минимального расстояния до границы – линяя линия левее зеленой.

3 Результат применения

Можно исследовать, как оценка погрешности зависит от количества имеющихся экспериментальных данных в «усредненном» случае, когда ответы нашей программы расположены в интервале на равных промежутках.

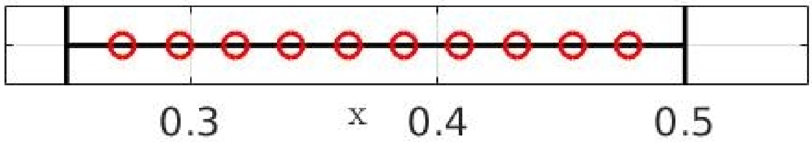

Рис. 3: Равномерное расположение 10 пробных точек в интервале $[0.25; 0.5]$.

На глаз зависимость на рис.(4) близка к $1/N$, что ожидаемо, т.к. погрешность в основном определяется минимальным расстояние до границы, которое при выбранной расстановке точек убывает как $1/N$.

Можно проверить отклонения от закона $1/N$ – рис.(5).

Видно, что наклон с правда близок к -1, но небольшие отклонения от линейности есть.

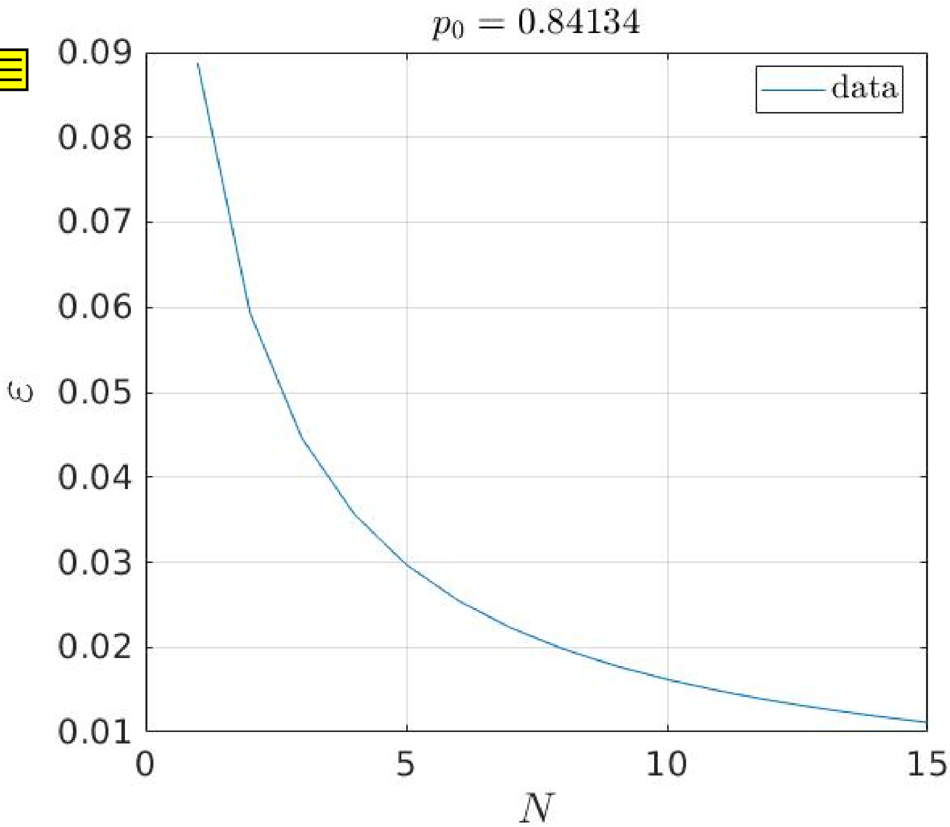

Рис. 4: $\varepsilon(N)$

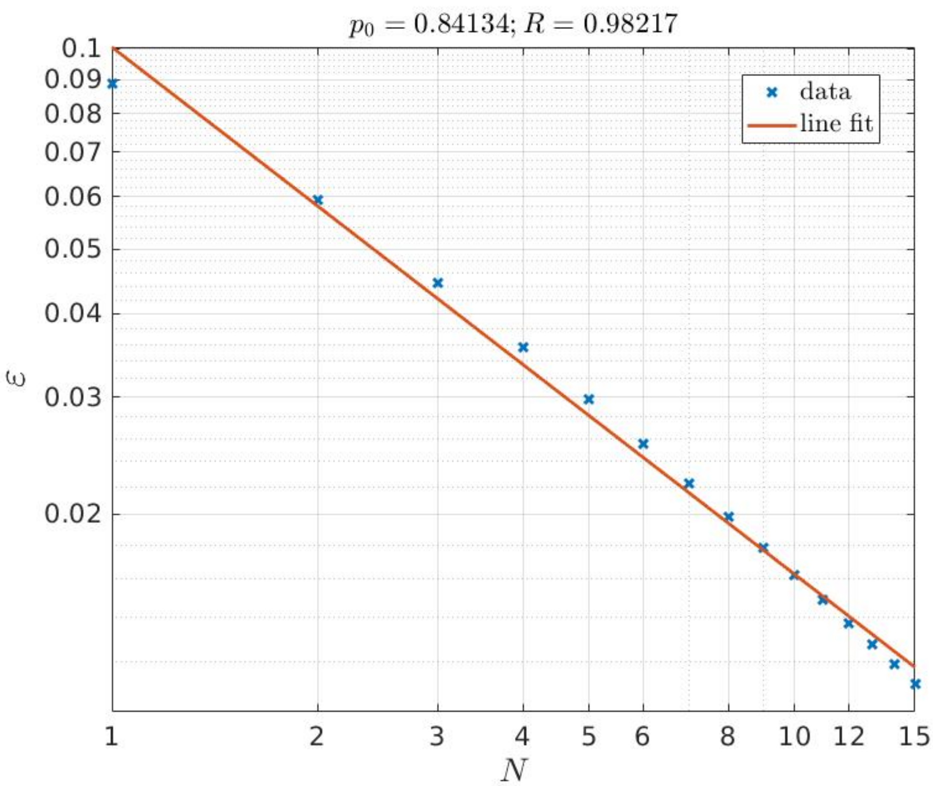

Рис. 5: $\varepsilon(N)$, логарифмический масштаб, попытка линеаризации