17 Серийное производство

Задача. Трент наладил серийное производство шифровальных машин `Amgine` и выпустил первую серию из n машин. Производственные ресурсы Трента практически неограниченны и Виктор не может сделать никаких априорных выводов об объеме серии. Однако известно, что машины снабжены последовательными серийными номерами от 1 до n и выдаются абонентам в случайном порядке. Виктор узнал, что Алиса получила машину с номером 539, Боб — с номером 734, а Глеб — с номером 222. Помогите Виктору оценить n.

Решение. Задача состоит в статистическом оценивании и, таким образом, не имеет однозначного решения. Рассмотрим различные оценки объема серии, которые могут пригодиться Виктору.

Будем анализировать общий случай, когда пользователи получают машины с номерами $m_1, m_2, \dots, m_k$. Пусть $m = \max m_i$. Обозначим $n^{[k]} = n(n-1)\dots(n-k+1)$ и пусть символ $\propto$ означает «пропорционально».

1. *Байесовская оценка.* Предположим, что объем серии — случайная величина с равномерным распределением на множестве $\{1, 2, \dots, N\}$, т. е. Трент произведет n машин с вероятностью

$$\pi(n) = \frac{1}{N}.$$

Здесь N — максимальные производственные ресурсы Трента (уточним позднее).

Условная вероятность того, что пользователи получат машины с номерами $m_1, m_2, \dots, m_k$ при объеме серии n есть

$$p(m_1, m_2, \dots, m_k \mid n) = \begin{cases} 1/n^{[k]}, & m \le n \le N, \\ 0, & \text{в противном случае.} \end{cases}$$

По формуле Байеса апостериорная вероятность

$$p(n \mid m_1, \dots, m_k) = \frac{p(m_1, \dots, m_k \mid n)\pi(n)}{p(m_1, \dots, m_k)} \propto 1/n^{[k]} \quad (m \le n \le N).$$

Адекватной оценкой n является медиана распределения $p(n \mid m_1, \dots, m_k)$, т.е. такое n_1, что суммы

$$\sum_{n=m}^{n_1} 1/n^{[k]} \text{ и } \sum_{n=n_1}^{N} 1/n^{[k]}$$

максимально близки.

Предполагая $k \ge 2$, считая $n^{[k]} \approx n^k$, заменяя суммы интегралами и устремляя N к бесконечности (ресурсы Трента неограниченны), получаем условие

$$1/m^{k-1} - 1/n_1^{k-1} \approx 1/n_1^{k-1},$$

откуда

$$n_1 \approx m \sqrt[k-1]{2} \approx m \frac{k}{k-1}.$$

В нашем случае $n_1 \approx 734\sqrt{2} \approx 1038$.

2. *Оценка Джеффриса.* В байесовском оценивании выбор априорного распределения является своего рода искусством. Для оценивания положительных параметров в условиях полной априорной неопределенности английский математик Джеффрис (Jeffreys) в своей книге [Theory of Probability. Oxford University Press, 1961] предложил использовать функцию

$$\pi(n) \propto \frac{1}{n}.$$

Оговоримся, что $\pi(n)$ не задает распределение вероятностей, поскольку ряд $\sum_{n=1}^{\infty} 1/n$ расходится. Тем не менее, статистические выводы на основе данной априорной функции оказываются адекватными.

Джеффрис использовал $\pi(n)$ при решении задачи о трамвае. Задача соответствует случаю $k = 1$ и звучит так: «Путешественник приехал в незнакомый город и увидел травмай номер m_1. Сколько всего в городе номеров трамваев?»

С новым априорным распределением

$$p(n \mid m_1, \ldots, m_k) \propto 1/(n \cdot n^{[k]}) \quad (m \le n \le N).$$

Для медианы этого апостериорного распределения справедливо приближение

$$n_2 \approx m\sqrt[k]{2} \approx m\frac{k+1}{k}.$$

В нашем случае $n_2 \approx 734\sqrt[3]{2} \approx 925$.

Интересно, что похожая оценка

$$n_2^* = (m-1)\frac{k+1}{k}$$

использовалась англичанами во время второй мировой войны для оценки числа сошедших с конвейера немецких танков T-V «Пантера». Английские статистики оценивали объем серии по номерам танков $m_1, m_2, \ldots, m_k$, захваченных на полях сражений (см. `http://www.guardian.co.uk/world/2006/jul/20/secondworldwar.tvandradio`).

3. *Метод моментов*. Наш читатель `CryptoOfficer` предложил воспользоваться следующим соображением: при объеме серии n среднее номеров выданных машин должно быть близко к $(n+1)/2$. Отсюда получаем оценку

$$n_3 = \frac{2}{k}(m_1 + m_2 + \ldots + m_k) - 1.$$

В нашем случае $n_3 \approx 996$.

На самом деле. На самом деле Трент произвел $n = 1000$ машин и выбирал случайные номера, используя знаки после запятой в разложении $e\pi = 8.539734222\ldots$.