Нормальное распределение

Кордзахия, Никулина, Скворцов & Шапаринов

Сквозь тернии к звездам

НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ обладает самой красивой функцией плотности среди всех распределений:

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

Но почему оно выглядит именно так, а не иначе? Вид нормального распределения — результат работы нескольких поколений ученых XVI-XIX столетий. Исток нормального распределения, как и исток многих человеческих достижений, кроется в человеческих ошибках — а точнее, в ошибках астрономических измерений.

Астрономические исследования требуют точности и в то же время подвержены ошибкам наблюдения. Вопрос о том, как получать максимально точные оценки координат небесных тел, был актуален еще со времен Древней Греции. Во втором веке до н. э. Гиппарх использовал среднее минимума и максимума; четыре же столетия спустя Птолемей грешил выбором наиболее удобного для объяснения значения. Ученые эпохи Возрождения привыкли к множественному сбору данных, но не смогли определиться с конвертацией их в единый показатель: помимо среднего и медианы, использовались уникальные и не всегда очевидные методы. Например, основываясь на следующих наблюдениях:

Прямое восхождение Марса 13/23 января 1600 г. в 11 часов 50 минут:

	°	′	″
используя яркую звезду стопы Близнецов	134	23	39
используя Регул	134	27	37
используя Поллукс	134	23	18
в 12 ч. 17 м., используя третью звезду крыла Девы	134	29	48
Среднее с учетом однозначности наблюдений:	134	24	33

Иоганн Кеплер получил собственную оценку прямого восхождения Марса — 134° 26′ 5.5″. До сих пор неизвестно, каким образом эта оценка была получена.

Как бы то ни было, астрономы того времени знали о существовании ошибок наблюдения и пытались с ними бороться.

Гауссова кривая

Впервые свойства случайных ошибок были описаны Галилео Галилеем в 1632 году, каждое из них до сих пор используется в рамках статистики и эконометрики:

- Существует одно, истинное значение наблюдаемой переменной.
- Все наблюдения подвержены ошибкам, связанным с наблюдателем, инструментарием и прочими условиями.
- Наблюдения распределены равномерно около истинного значения. Значит, ошибки распределены равномерно вокруг нуля.
- Маленькие ошибки более вероятны, нежели большие.

Из утверждений Галилео следовало, что при достаточно большой выборке для оценки истинного значения достаточно посчитать среднее всех наблюдений, так как даже серьезные отклонения будут нивелированы своей малой частотой появления и не вызовут сильных искажений.

Но как формализовать функцию таким образом, чтобы минимизировать вероятность сделанной ошибки? Ответ на этот вопрос был дан почти 200 лет спустя; в 1801 году итальянский астроном Джузеппе Пьяцци обнаружил небесное тело, ныне именуемое Церерой, и сделал предположение о ее планетарном характере. К сожалению, сделанных наблюдений не хватило, чтобы надежно вычислить орбиту Цереры: через полгода после обнаружения она скрылась за Солнцем. Перед астрономами встала задача вычислить место ее повторного появления, основываясь на 21 наблюдении, сделанном Пьяцци.

Для вычисления места второго появления Цереры Гаусс использовал разработанный им метод наименьших квадратов, который утверждает, что лучшей оценкой некоего истинного параметра μ является среднее его наблюдений, поскольку оно решает задачу минимизации квадратичной функции ошибок:

$$\arg\min_{\mu} \sum \epsilon_i^2 = \sum (y_i - \mu)^2 = \frac{y_1 + \ldots + y_n}{n} = \bar{y}$$

Чтобы защитить этот подход, Гауссу необходимо было найти такое распределение ошибок, которое согласовывалось бы с предпосылками, сделанными Галилео, и обеспечивало бы максимальную правдивость имеющихся сведений при $\hat{\mu} = \bar{y}$.

Таким образом, **Гаусс делает следующие предположения:**

- μ — истинное значение;
- $y_i = \mu + \epsilon_i$, где ϵ_i — ошибка измерения;
- ϵ_i — н.о.р.с.в., симметрично распределенные относительно нуля;
- Вероятность правдивости имеющихся сведений максимальна, если за истинное значение принимается среднее: $\arg\max_{\mu} f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \mu) = \bar{y}$.

Оказалось, что этим требованиям удовлетворяет именно колоколообразная кривая!

Доказательство

Решим задачу по поиску максимума функции плотности этого случайного вектора:

$$L = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \mu) = f_{Y_1}(y_1; \mu) \cdot f_{Y_2}(y_2; \mu) \cdot \ldots \cdot f_{Y_n}(y_n; \mu)$$

При этом заметим, что $f_{Y_i}(y_i; \mu) = f_{\epsilon_i}(y_i - \mu)$. Тогда

$$L = f_\epsilon(y_1 - \mu) \times f_\epsilon(y_2 - \mu) \times \ldots \times f_\epsilon(y_n - \mu)$$

Проведем монотонное преобразование:

$$\ln L = \ln f_\epsilon(y_1 - \mu) + \ldots + \ln f_\epsilon(y_n - \mu) \to \underset{\mu}{max}$$

$$\frac{\partial \ln L}{\partial \mu} = \frac{f'_\epsilon(y_1 - \mu)(-1)}{f_\epsilon(y_1 - \mu)} + \ldots + \frac{f'_\epsilon(y_n - \mu)(-1)}{f_\epsilon(y_n - \mu)} = 0$$

$$\sum_i \frac{f'_\epsilon(y_i - \mu)}{f_\epsilon(y_i - \mu)} = 0$$

помню, что делалось, поэтому надо разбираться]

Всего $n+1$ свидетелей. Предположим, что n говорят одно то же и некий Вася дает совсем другой ответ. Например

$$y_1 = 4, y_2 = 4, y_3 = 4, y_4 = 4, y_5 = 9 \Rightarrow \bar{y} = 5$$

$$\Rightarrow \sum_i = n\frac{f'_\epsilon(\alpha)}{f_\epsilon(\alpha)} + \frac{f'_\epsilon(-n\alpha)}{f_\epsilon(-n\alpha)} = 0$$

$$n\frac{f'_\epsilon(\alpha)}{f_\epsilon(\alpha)} = \frac{f'_\epsilon(n\alpha)}{f_\epsilon(n\alpha)}$$

$$\frac{f'_\epsilon(\alpha)}{f_\epsilon(\alpha)} = \frac{1}{n}\frac{f'_\epsilon(n\alpha)}{f_\epsilon(n\alpha)} \qquad (1)$$

На основании формулы (1) ББ, насколько я помню, заявил о линейности функции f

$$\frac{f'_\epsilon(\alpha)}{f_\epsilon(\alpha)} = \gamma\alpha$$

$$\frac{\mathrm{d}f/\mathrm{d}\alpha}{f} = \gamma\alpha$$

$$\int \frac{\mathrm{d}f}{f} = \int \gamma\alpha \mathrm{d}\alpha$$

$$\ln f = \frac{1}{2}\gamma\alpha^2 + C$$

$$f = C \times \exp(\frac{1}{2}\gamma\alpha^2), \gamma < 0, C > 0$$

Пусть $\gamma = -\frac{1}{\sigma^2}$

$$f_\epsilon(x) = C \exp(-\frac{x^2}{2\sigma^2}) \to \epsilon_i \sim N(0, \sigma^2)$$

$$\mu + \epsilon_i = Y_i \sim N(\mu, \sigma^2)$$

LaTeX TikZposter