Научно-учебная группа

Исследование частотных характеристик языка

Санкт-Петербург
2023

Распределения Ципфа-Парето-Мандельброта и их особенности

Спикер: Крайторов Михаил Владимирович
Модераторы: Горина Ольга Григорьевна и Рагозин Илья Андреевич

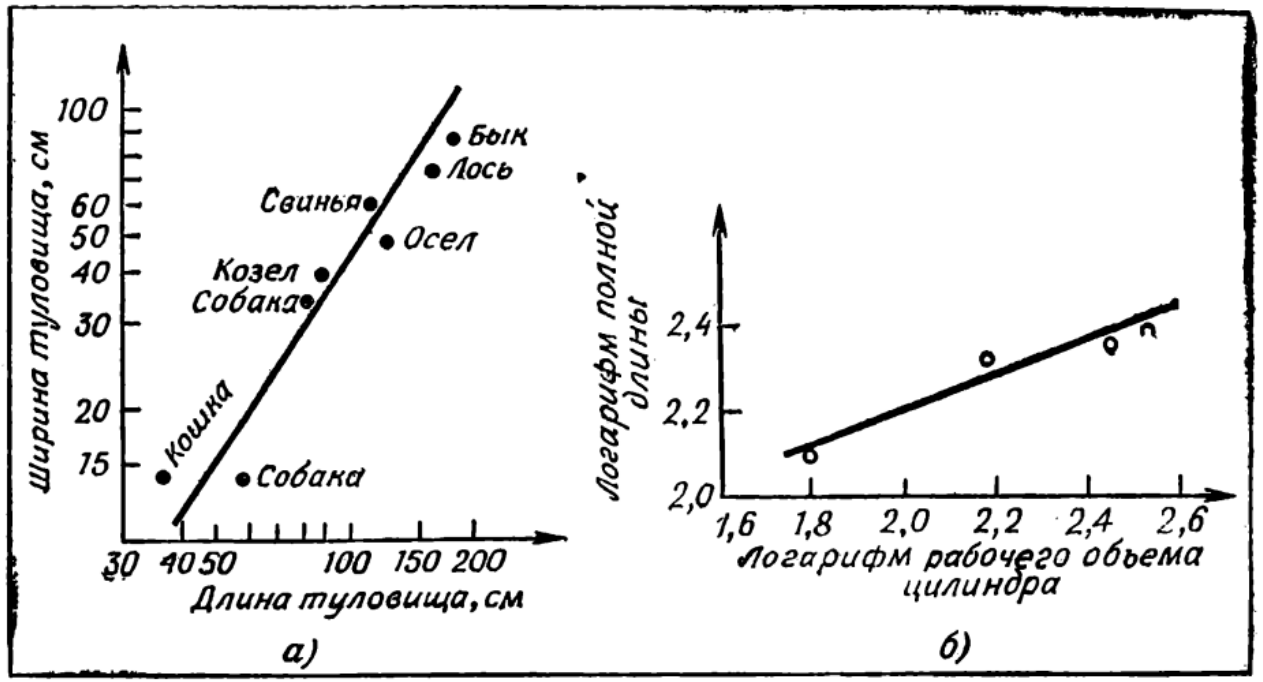
Введение в гиперболические распределения

В современных науках, связанных с изучением биологических, технологических, социально-экономических и информационных систем, для многих исследуемых параметров весьма распространена функциональная зависимость в виде пропорционального соотношения между одной переменной y и другой переменной x, возведенной в ту или иную степень.

Называемая по этой причине степенной, эта зависимость имеет следующий вид, где A и α константы.

$$Y = Ax^{\alpha}$$

Введение в гиперболические распределения



Примеры аллометрического закона. Р. Розен.
Принцип оптимальности в биологии. М., «Мир», 1969.

Введение в гиперболические распределения

Зависимости этого типа часто называются гиперболическими, так как произведение координат обычной гиперболы относительно ее асимптот также равно константе. Величины соответствующих параметров находятся в обратном соотношении между собой; в частности, при γ = 1 они просто обратно пропорциональны.

$$Y = \frac{A}{x^{\gamma}}$$

Введение в гиперболические распределения

Гиперболические закономерности широко распространены в биологических, социальных, информационных процессах и подтверждаются обширным статистическим материалом.

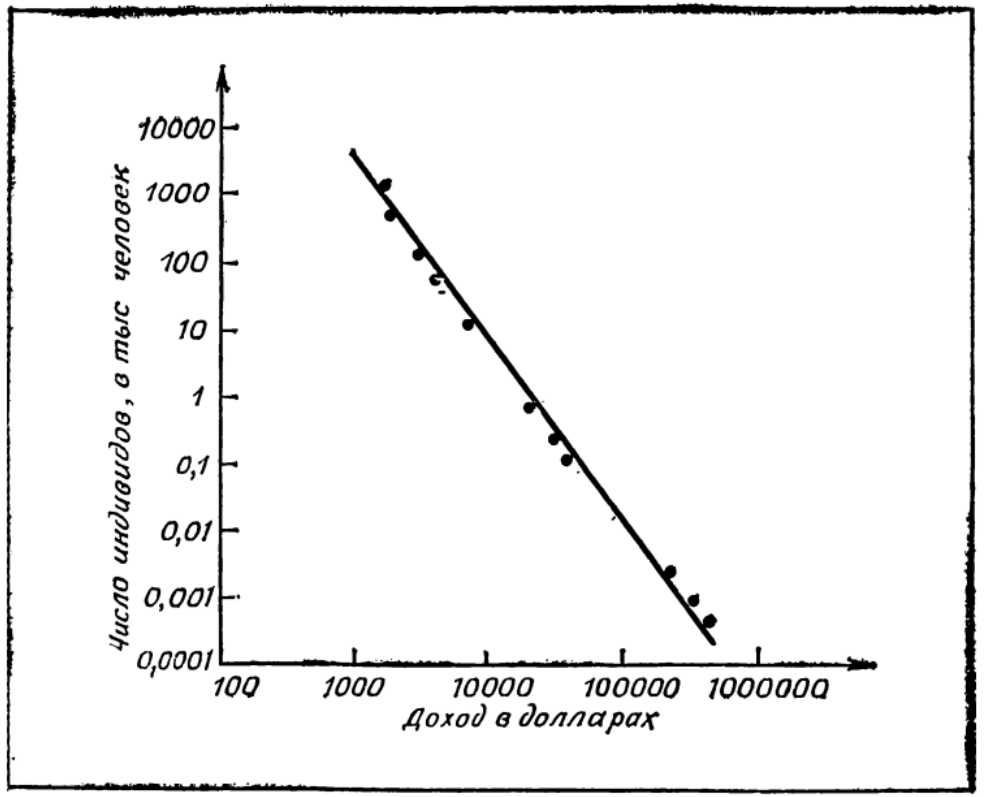
Один из наиболее распространенных методов обработки эмпирических данных сводится к тому, что для ограниченного множества подсчитывается число элементов, обладающих данным значением параметра x, а затем устанавливается распределение этого числа элементов n(x) в зависимости от величины соответствующего параметра x. Оказывается, что при достаточно большом числе элементов имеет место распределение:

$$n(x) = \frac{A}{x^{1+\alpha}}$$

Распределение доходов в США (1918 г.)

Один из первых эмпирических результатов, описываемых гиперболическим распределением, был получен около ста лет назад известным итальянским экономистом Вильфредо Парето.

Он вывел кривую распределения доходов на базе статистических данных о подоходном налоге. Эта кривая подтверждается в соответствующих пределах и для современного распределения доходов в капиталистических странах.



Логарифмическая шкала по обеим кординатам.

E. W. Montroll, W. W. Badger. Introduction to Quantitative Aspects of Social Phenomena, Gordon and Breach Science Publishers, New-York—London—Paris 1974.

Частотный подход

Во всех рассмотренных случаях задача сводилась к тому, чтобы **подсчитать число элементов** (людей с теми или иными доходами, «биологических родов,ученых и пр.), **связанных с соответствующим значением некоторого параметра** (величины дохода, числа биологических видов, числа публикаций и пр.), **и определить зависимость числа элементов от величины характеризующего их параметра**.

При соответствующей нормировке иногда удобнее говорить не об абсолютном числе n(x) таких элементов, а об их доле p(x) во всем анализируемом массиве, или о **частоте** встречаемости элементов с данным значением параметра.

α здесь является некоторой мерой неравенства в распределении параметра. Возрастание α приводит к увеличению вогнутости кривой соответствующего гиперболического распределения (к увеличению разрыва между высокопродуктивными и малопродуктивными учеными, между людьми с большими и малыми доходами и пр.).

$$n(x) = \frac{A}{x^{1+\alpha}}$$

Ранговый подоход

Для множества **ранжированных** по уменьшению параметра элементов, составляющих информационную или социально-экономическую систему (например, литературный текст, состоящий из слов с разной частотой повторяемости; географический регион, состоящий из городов с разным населением, и пр.), во многих случаях имеет место гиперболическая зависимость следующего вида, где r - ранг элемента; B, β - параметры.

*Эта зависимость носит название **ранговой**, а сам метод ее определения называется **ранговым подходом** к эмпирическому анализу гиперболических распределений.*

$$x(r) = \frac{B}{r^{\beta}}$$

История рангового подхода

Эмпирический анализ гиперболических распределений ранговым методом был впервые предложен Дж. Ципфом, исследовавшим лингвистический материал. В рамках данного метода для конкретно взятого текста, последовательность всех различных слов ранжируется в порядке убывания частотности и сопоставляется с его местом, или рангом.

Наряду с законом Ципфа примером рангового подхода является эмпирическое распределение, прослеживаемое при анализе массива научных журналов, которое называется обычно **законом Брэдфорда**, или законом рассеивания научной информации. Этот закон называют иногда, в силу его важности, **основным библиометрическим законом** или **основным законом информации**.

Стремясь найти закономерность, которой подчиняется распределение научной информации по данной тематике в различных научных журналах, Брэдфорд отобрал журналы, в которых содержались одна, две и более статей на одну определенную тему (в его исследовании это были прикладная геофизика и одна из областей технологии), и статистически обработал этот информационный массив.

Оказалось, что последовательность множества журналов, ранжированных в порядке уменьшения числа статей по данной тематике (от наиболее продуктивного журнала с максимальным числом статей и рангом, равным единице, до наименее продуктивного журнала с одной статьей, замыкающего ряд), можно разделить на группы с приблизительно одинаковым суммарным числом статей в каждой группе.

Закон Брэдфорда

В математическом плане закон Брэдфорда в простейшей формулировке утверждает, что общее число статей по данной тематике в первых n наиболее продуктивных журналах пропорционально логарифму от числа n этих журналов.

Проинтегрировав закон Ципфа (принимаем, что r меняется непрерывно), мы получаем закон Брэдфорда.

$$x = \frac{B}{r},\ B - const$$

$$R(n) = \int_1^n \frac{B}{r} dr = B\ ln\ n,\ n_0 = 1,\ B - const$$

Взимосвязь частотного и раногового методов

Пусть имеется совокупность элементов определенного типа. Каждый из элементов снабжен меткой, выбираемой из некоторого множества.

Пусть n(x) — число различных меток, каждая из которых встречается ровно x раз в данной совокупности элементов.

Тогда для достаточно большой совокупности элементов имеет место эмпирическая зависимость, аналогичная рассмотренному выше частотному представлению гиперболического распределения.

$$n(x) = \frac{A}{x^{\gamma}} = \frac{A}{x^{1+\alpha}};\ \gamma = 1 + \alpha$$

Для числа меток r (x), встречающихся x раз и более, с учетом формулы выше получаем следующее выражение (сумму заменим интегралом):

$$r(x) = \sum_{\xi=x}^{\infty} n(\xi) \approx \frac{A}{\alpha} \cdot \frac{1}{x^{\alpha}} = \frac{C}{x^{\alpha}};\ C = \frac{A}{\alpha}$$

Взимосвязь частотного и раногового методов

Если все метки расположены в ряд в порядке убывания их встречаемости (уменьшения x), т. е. проранжированы, то величина r, называемая рангом и определяемая выражением предыдущим, есть положение в этом ряду метки, встречающейся x раз (порядковый номер этой метки).

Меняя в выражении выше x и r местами, мы и переходим от частотного к аналогичному ранговому представлению гиперболического распределения, устанавливая тем самым взаимосвязь между этими подходами.

$$x(r) = \frac{C^{1/\alpha}}{r^{1/\alpha}} = \frac{B}{r^{\beta}};$$

$$B = C^{1/\alpha},\ \beta = \frac{1}{\alpha}$$

Особенности Ципфа-Парето распределений

Одним из основных эмпирических фактов, противоречащих гауссовскому представлению в приложении к закону Ципфа — Парето, является эффект концентрации соответствующих параметров на «слишком» малом (по сравнению с гауссовским характером случайной выборки) числе элементов статистического массива, описываемого этим законом. Этот эффект выражается, например, в том, что около 5 % наиболее продуктивных журналов могут содержать до 70 % всех статей по данной тематике, преобладающая часть городского населения сосредоточена в крайне небольшом числе больших городов, при общем числе 100 авторов около 10 высокопродуктивных пишут до половины всего массива статей и т. д.

Почему Ципфа-Парето распределение, а не гауссовское?

В законе Ципфа — Парето существуют моменты только порядка $k < \alpha$, где α — характеристический показатель. Экспериментальные данные показывают, что в большинстве практических приложений закона Ципфа — Парето $\alpha < 2$, а это по определению приводит к бесконечной дисперсии. В таком случае сходимость закона Ципфа — Парето к закону Гаусса не имеет места, так как условием сходимости к нему по центральной предельной теореме является конечность второго момента.

Таким образом, негауссовский характер закона Ципфа— Парето заставляет отказаться от гауссовского представления по отношению к закону Ципфа — Парето и считать, что в его основе лежат принципиально иные, «негауссовские» закономерности. В современной теории вероятностей наличие таких закономерностей предусмотрено и даже создана математическая теория для их исследования. Эта теория известна под названием теории устойчивых распределений.

Почему Ципфа-Парето распределение, а не гауссовское?

Стоит отметить, что в статье “СТАТИСТИКА ЦИПФА-ПАРЕТО-МАНДЕЛЬБРОТА И АНАЛИЗ ПАРЕТО” (С.А. Щеглова, 2002) показано, что известный метод кумулянт Парето базируется на "патологически" негауссовских статистиках. В их основе лежат статистические ансамбли, в которых действуют механизмы конкуренции. Они подчиняются принципу неэквивалентного микрообмена в процессах перемешивания. Авторами был сформулирован принцип, приводящий к классу распределений Ципфа-Парето-Мандельброта. А статистики Ципфа-Парето-Мандельброта, в свою очередь, возникают в таких сложных системах, где удается достигнуть компромисса между мерой возможностей – мерой неопределенности и сложности систем и логарифмическими затратами.


ВЫСШАЯ • ШКОЛА • ЭКОНОМИКИ •