Измерение объективной степени случайности конечного набора точек

В.И. Арнольд

июнь 2009

1. Определение двух параметров стохастичности

Вопрос о том, насколько “объективно случайна” наблюденная выборка значений — один из труднейших в прикладной математике. Описанные ниже числовые характеристики стохастичности, β и λ, доставляют некоторые объективные меры степени случайности данного конечного набора вещественных чисел (или точек окружности). Объективны они в том смысле, что выводы не зависят от происхождения исследуемого набора (который может оказаться “случайным” и доставляясь простым алгоритмом, вроде “датчика случайных чисел” теоретико-числового происхождения).

Хотя оба эти параметра, β и λ, кажутся совершенно независимыми характеристиками выборки, описанные ниже эксперименты показывают, что между их усредненными вариантами $\hat{\hat{\beta}}$ и $\hat{\hat{\lambda}}$ появляется неожиданная эмпирическая функциональная зависимость (16).

Мы будем рассматривать системы (“выборки”) k точек *конечной* (*дискретной*) *окружности* $\mathbb{Z}_n = \mathbb{Z}/(n\mathbb{Z})$ длины n.

Определение 1. (см.[1]). *Параметр стохастичности* B принимает на данной выборке значение

$$B = a_1^2 + a_2^2 + ... + a_k^2, \tag{1}$$

где $a_j \geqslant 0$ — длины тех k дуг, на которые изучаемые k точек делят окружность, $\mathbb{Z}_n$ длины n.

$$a_1 + a_2 + \ldots + a_k = n. \tag{2}$$

Наименьшее значение B_0 параметра стохастичности B достигается в центре симплекса (2). Оно равно

$$B_0 = k(n/k)^2 = n^2/k.$$

Среднее значение B_* параметра B примерно вдвое больше минимального: в [1] доказано, что

$$B_* = B_0(2k/(k+1)). \tag{3}$$

Определение 2. (см.[1]). *Безразмерным параметром стохастичности* (системы k точек дискретной окружности $\mathbb{Z}_n$) называется значение β параметра стохастичности B (определенного формулой (1)) нормализованное его делением на наименьшее возможное значение B_0 параметра B :

$$\beta = B/B_0 \tag{4}$$

Нормализующее деление на B_0 позволяет сравнивать между собой значения параметров стохастичности для окружностей разных длин n (в то время, как значение исходного параметра B имеет размерность площади, будучи пропорциональным квадрату длины рассматриваемой окружности).

Замечание. По определению, $\beta \geqslant 1$, а в среднем $\beta \approx \beta_*$, (где $\beta_* = 2k/(k+1) \approx 2$ при больших k по формуле (3)). Меньшие среднего значения (грубо говоря, $\beta < 2$) соответствуют взаимному *отталкиванию* между соседними точками (которое в предельном случае $\beta = 1$ расставляет их "военным" строем со всюду равными n/k расстояниями между соседними точками "*кристаллической решетки*" точек).

Бóльшие среднего значения (грубо говоря, $\beta > 2$) соответствуют взаимному *притяжению* соседних точек (приводящему к их скучиванию в "кластеры", вплоть до предельного случая максимального значения

$$\beta_1 = k, \tag{5}$$

реализуемого вершинами симплекса (2), т.е. скоплением всех k точек в одном месте: $a_1 = a_2 = \ldots = a_k = 0$).

Оба эти случая — отталкивания и притяжения — означают, что изученные точки *зависимы* — степень их взаимной зависимости измеряется разностью $\beta - \beta_*$.

Если изучаемые k точек независимо друг от друга брошены на окружность $\mathbb{Z}_n$, то вычисленное по ним значение безразмерного параметра стохастичности β не слишком далеко от среднего значения β_*, т.е. (при больших k) должно наблюдаться $\beta \approx 2$.

Совершенно другой параметр стохастичности, λ, был предложен А.Н. Колмогоровым [2] в итальянской статье 1933 года. В применении к нашей ситуации (k точек на дискретной окружности $\mathbb{Z}_n$ длины n) значение этого параметра вычисляется следующим образом.

Обозначим наши k точек значениями целочисленных остатков от деления их координат на n:

$$0 \leqslant x_1 \leqslant x_2 \leqslant \ldots \leqslant x_k \leqslant n. \tag{6}$$

Вместо выборки остатков (6) А.Н. Колмогоров рассматривает *считающую функцию* C_k со значениями

$$C_k(\Lambda) = (\text{число значений } x_j \leqslant \Lambda). \tag{7}$$

Эта считающая функция вещественного аргумента Λ(где $0 \leqslant \Lambda \leqslant n$) кусочно-постоянная между изучаемыми точками: если $x_1 < x_2 < \ldots < x_k$, то

$$C_k(\Lambda) = 0, \text{ при } \quad \Lambda < x_1,$$

$$C_k(\Lambda) = 1, \text{ при } \quad x_1 \leqslant \Lambda < x_2,$$

$$C_k(\Lambda) = m, \text{ при } \quad x_m \leqslant \Lambda < x_{m+1},$$

$$C_k(\Lambda) = k, \text{ при } \quad \Lambda \geqslant x_k \text{ (например, } C_k(n) = k).$$

Предполагая, что выборка (6) есть (записанный в возрастающем порядке) набор независимых значений одной и той же случайной величины, Колмогоров рассматривает "теоретически считающую функцию"

$$C_0(\Lambda) = (\text{математическое ожидание числа } x_m < \Lambda) = k(\text{вероятность события } x < \Lambda). \qquad (8)$$

Это — тоже неубывающая функция, тоже растущая от $C_0(0) = 0$ до $C_0(n) = k$. В среднем функции C_k и C_0 растут одинаково (прирастая на k при увеличении Λ на n).

Здесь Колмогоров предполагает, что функция C_0 непрерывна. В этом случае он доказывает поразительную (релятивистскую) теорему об универсальном распределении для отклонения C_k от C_0 при больших k.

А именно, отклонение S_k определяется у него формулой

$$S_k[x] = \sup_\Lambda |C_k(\Lambda) - C_0(\Lambda)|. \qquad (9)$$

Супремум стоит здесь вместо максимума потому, что считающая функция C_k имеет разрывы в точках x_m.

Релятивистской эта теория является в том смысле, что расстояние (9) не зависит от выбора координаты на оси Λ: выбор Λ^2 привёл бы к тому же значению расстояния S.

Отклонение $C_k - C_0$ можно истолковать как сумму k слагаемых разностей с математическим ожиданием нуль у каждого из них (поскольку в среднем C_k и C_0 растут одинаково). Поэтому естественно ожидать, что эта сумма — величина порядка $\sqrt{k}$ (как для суммы независимых слагаемых).

Поэтому, чтобы сравнивать выборки с разными значениями влияние k, Колмогоров делит значение Sна $\sqrt{k}$, определяя свой (нормализованный) параметр стохастичности λ_k формулой

$$\lambda_k[x] = S_k[x]/\sqrt{k}. \qquad (10)$$

Теорема Колмогорова состоит в том, что при $k \to \infty$ случайные величины λ_k сходятся (в смысле равномерной сходимости их функций распределения) к универсальной случайной величине Колмогорова, λ.

Универсальность состоит здесь в том, что функция распределения предельной величины λ не зависит от того, с какой (непрерывной) функции распределения C_0 начиналась конструкция.

А именно, Колмогоров вычисляет свою универсальную предельную функцию распределения Φ ($\Phi(\Lambda)$= вероятность события $\lambda \leqslant \Lambda$) как θ-функцию следующего вида:

$$\Phi(\Lambda) = \sum_{m=-\infty}^{\infty} ((-1)^m e^{-2m^2\Lambda^2}, \quad \Lambda \geqslant 0. \qquad (11)$$

Распределение Колмогорова ведет себя так:

$\Phi(0) = 0$ со всеми производными, а при $\Lambda > 0$ эта функция монотонно растет до $\Phi(\infty) = 1$.

При этом, например, Колмогоров вычислил значения

$$\Phi(0,4) \approx 0,003, \quad \Phi(1,8) \approx 0,997.$$

Можно вычислить среднее по распределению Колмогорова Φ значение его параметра Λ оно составляет $\Lambda_* = \sqrt{\frac{\pi}{2}}\ln 2 \approx 0,87$

Как слишком большие, так и слишком малые значения параметра стохастичности Колмогорова данной выборки x указывают на весьма малую вероятность случайности этой выборки (как событие $\lambda \leqslant 0,4$, так и событие $\lambda \geqslant 1,8$, имеют вероятности меньше трети процента).

Пример. Рассмотрим следующие две последовательности, каждая из которых состоит из 15 двузначных чисел (остатков от деления на 100).

$$03, 09, 27, 81, 43, 29, 87, 61, 83, 49, 47, 41, 23, 69, 07 \tag{12}$$

$$37, 74, 11, 48, 85, 22, 59, 96, 33, 70, 07, 44, 81, 18, 55 \tag{13}$$

На вид обе последовательности одинаково случайны, но вычисление значений параметра Колмогорова показывают, что первая последовательность (геометрическая прогрессия) примерно в 300 раз случайнее второй (арифметической прогрессии).

А именно, (предполагая теоретическое распределение равномерным) формулы (9), (10) дают соответственно:

$$\lambda_{15}^{(12)} \approx 0,70, \quad \lambda_{15}^{(13)} \approx 0,33.$$

Из формулы (11) в этих точках получаются значения:

$$\Phi(0,70) \approx 0,30, \quad \Phi(0,33) \approx 0,001.$$

Вероятность столь малого значения $\lambda \approx 0,33$, какое доставляет последовательность (13), примерно в 300 раз меньше вероятности (умеренного) значения λ порядка 0,77 (близкого к Λ_*).

Применять в нашем случае (точек на конечной окружности $\mathbb{Z}_n$) теорему Колмогорова *формально* нельзя, так как его предположения не выполняются.

Во-первых, дискретное распределение (точек из $\mathbb{Z}_n$) не может иметь непрерывной функции распределения.

Во-вторых, топология окружности (гомологии) отлична от топологии прямой R, на которой Колмогоров рассматривает свои считающие функции.

Поэтому я предлагаю обобщить теорему Колмогорова на случай случайных величин со значениями на окружности следующим образом.

Считающие функции — это, в сущности, *первообразные* изучаемых распределений точек. На прямой Колмогоров выбирает в качестве первообразной определенный интеграл

$$I(\Lambda) = \int_{-\infty}^{\Lambda} f(\lambda) d\lambda.$$

На окружности начала отсчета "$-\infty$" нет, так что первообразных много. Первообразная определена с точностью до выбора в неопределенном интеграле "постоянной интегрирования".

Чтобы избежать этой топологической трудности, я предлагаю модифицировать формулу расстояния (9) теории Колмогорова следующим образом.

Во-первых, выберем обе сравниваемые первообразные (на накрывающей прямой) произвольно.

Во-вторых, вычтем из разности (9) постоянную так, чтобы *минимизировать* расстояние между первообразными: постоянные интегрирования надо выбрать так, чтобы минимизировать модуль разности.

Иными словами, формула (9) заменяется на формулу

$$S_k = \inf_c \sup_{\Lambda \in R} |C_k(\Lambda) - C_0(\Lambda) - c| \tag{14}$$

Определенное нижней гранью (14) расстояние уже не зависит от сделанного выбора первообразных.

Я предполагаю, что соответствующие варианты теоремы Колмогорова (об универсальности) должны выполняться и для случайных величин со значениеми на окружности (с расстоянием (14) вместо (9)), и для целочисленных случайных величин ($x \in \mathbb{Z}$ или $x \in \mathbb{Z}_n$) со значениями на целочисленной прямой или на целочисленной окружности.

При этом, если шаг дискретной решетки достаточно мал, то соответствующее гипотетическое универсальное предельное распределение будет предположительно близким к распределению Колмогорова.

Я не доказываю этих гипотез и не использую их ниже, но без них настоящая работа не имела бы быть придумана: она основана на физическом предположении, что дискретные распределения на решетках с достаточно малым шагом не должны давать других явлений, чем истинно непрерывные распределения.

Главное в теореме Колмогорова — релятивистское соображение, которое ее и доказывает: все распределения с непрерывными функциями распределения переводятся друг в друга сохраняющим меру гомоморфизмом, а потому все их инвариантные относительно этих преобразований свойства одинаковы.

2. Совместное распределение обоих параметров стохастичности

В качестве примера ниже рассматриваются случаи системы $k = 6$ точек на окружности $\mathbb{Z}_{17}$ длины $n = 17$.

Число таких конфигураций (рассматриваемых с точностью до 17 вращений окружности $\mathbb{Z}_{17}$) равно $728 (= 3^6 - 1)$ [1]. Для каждой из этих 728 конфигураций я сосчитал в [3] значения параметров стохастичности β и λ. Получилось 728 "хаотически распределенных" точек на плоскости с координатами β и λ.

Чтобы избавиться от от влияния локальных хаотических флуктуаций, я разбил все эти 728 точек на десяток групп близких точек и заменил каждую такую группу ее центром тяжести, $(\hat{\hat{\beta}}, \hat{\hat{\lambda}})$,

Эти центры тяжести оказались уже менее хаотическими, чем исходные 728 точек — десяток центров тяжести удивительно точно попадает на вполне регулярную кривую (а именно, на неожиданную параболу) на плоскости с координатами β и λ.

Разбиение на группы производилось так: в группу номер m включались те конфигурации, для которых значения размерного параметра стохастичности B лежат в интервале

$$10m \leqslant B < 10(m+1). \tag{15}$$

Средние значения величин ($\hat{\hat{B}}$, $\hat{\hat{\lambda}}$, $\hat{\hat{\beta}}$ и т.д.) определяются как средние арифметические значений соответствующих величин (B, λ, β и т.д.) для всех членов группы номер m.

В [3] доказывается

Теорема. *Средние значения по группам (15) имеют для 728 конфигураций $k = 6$ точек*

[1] Это видно, например, из того, что $C_{17}^6/17 = \frac{16 \cdot 15 \cdot 14 \cdot 13 \cdot 12}{720} = 27^2 - 1$.
Мы воспользовались тем, что никакое нетривиальное вращение окружности $\mathbb{Z}_{17}$ не переводит в себя никакое разбиение этой окружности на 6 дуг, поскольку число 17 — простое. Никакой связи шестерок точек из $\mathbb{Z}_{17}$ с полями Галуа из 3^6 элементов я не знаю.

конечной окружности Z_{17} следующие приближенные значения:

m	4	5	6	7	8	9	10	11	12	14
$\hat{\hat{B}}$	49	56,7	64,5	73,5	84	96	101	112	129	149
$\hat{\hat{\beta}}$	1,02	1,17	1,33	1,52	1,74	1,99	2,09	2,32	2,67	3,08
$1000\overleftarrow{\hat{\lambda}}$	264	387	471	530	591	651	681	725	791	865

Доказательство состоит в переборе 728 случаев, каждый из которых исследуется геометрически примерно за четверть часа. Получившиеся десять усредненных точек на плоскости с координатами $\hat{\hat{\lambda}}$ и $\hat{\hat{\beta}}$ удивительно близко лежат к параболе

$$\hat{\hat{\beta}} = p + q\hat{\hat{\lambda}} + r\hat{\hat{\lambda}}^2, \tag{16}$$

(где $p \approx 1,11$, $q \approx -1,57$, $r \approx 4,44$.)

Никакой причины для этой эмпирически найденной неожиданной функциональной зависимости я не знаю, но вот ее график.

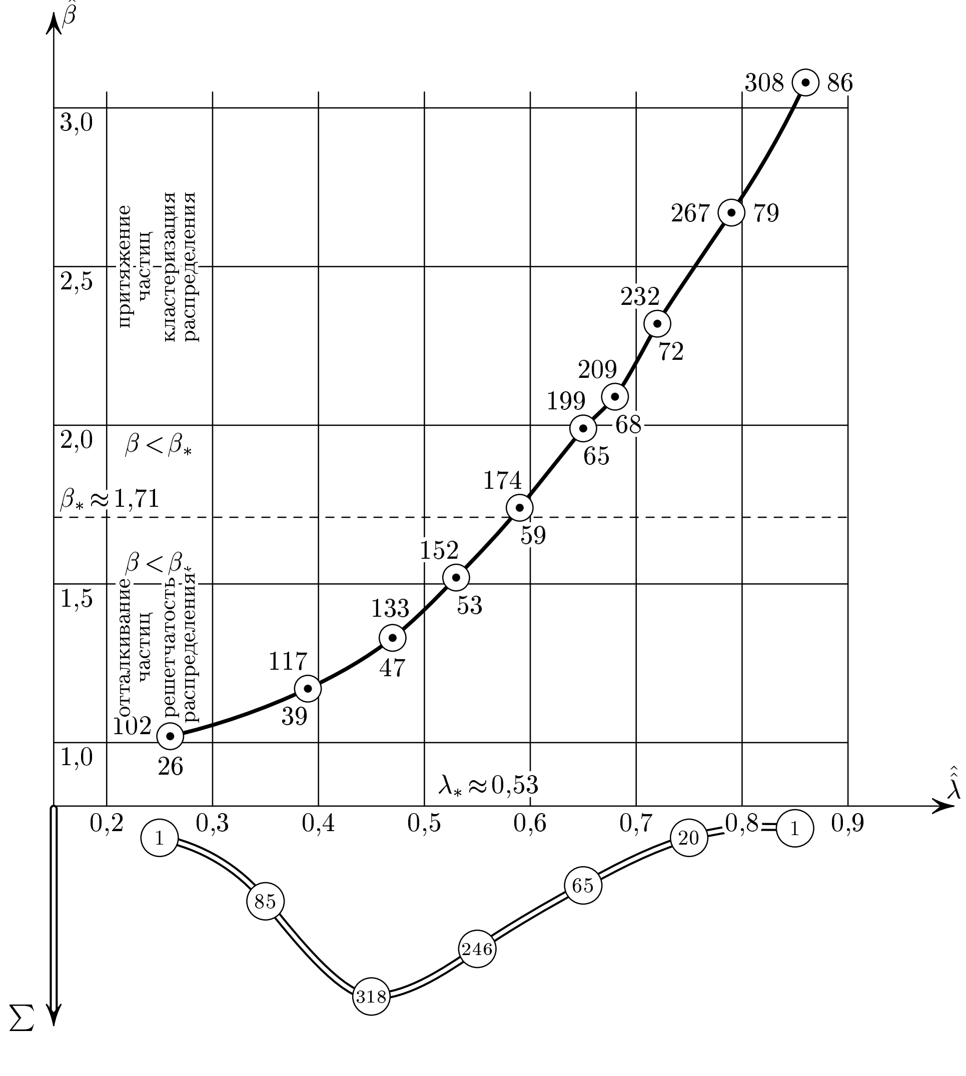
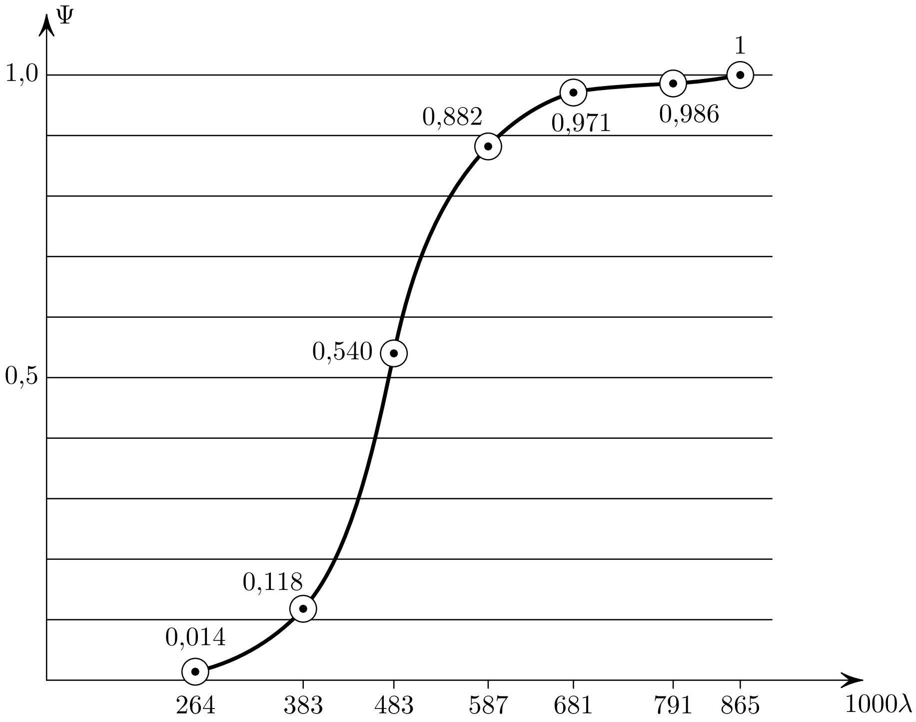


Рис. 1. Гистограмма распределения 728 значений параметра λ

Вычисления, доказывающие приведенную теорему, доставляют также негауссовскую функцию распределения Ψ значений параметра λ (для распределения $r = 6$ точек на окружности $\mathbb{Z}_{17}$):



В качестве среднего значения параметра стохастичности λ системы 6 случайных точек на окружности длины 17 это распределение доставляет значение $\lambda_* \approx 0,532$.

Замечание. Эти данные показывают следующее отличие предполагаемого универсального распределения Ψ параметра λ (для систем точек на окружности) от универсального распределения Φ теоремы Колмогорова: распределение в случае окружности, Ψ, систематически сдвинуто влево, в сторону меньших значений параметра стохастичности λ по сравнению с Φ. Например, среднее значение параметра стохастичности λ для распределения Колмогорова $\lambda_{*\Phi} = \sqrt{\pi/2}\ln 2 \approx 0,87$ примерно на 60% выше, чем среднее значение $\lambda_{*\Psi} \approx 0,53$, получающееся для случайных 6 точек на дискретной окружности длины 17.

Цитированная литература.

[1] В.И. Арнольд. Группы Эйлера и арифметика геометрических прогрессий. — МЦНМО, Москва,2003, 44 стр.

[2] A.N. Kolmogorov. Sulla determinazione emirica di una legge di distribuzione. — G. Ist. Ital. Attuari, 4 (1933), 83–91.

[3] V.I. Arnold. Comparison of two stochasticity parameters. — FAOM, Springer, 2009, 20 pp.

Резюме.

Для случайного распределения k точек на целочисленной окружности $\mathbb{Z}_n$ длины n два "параметра стохастичности" β и λ были определены (независимо друг от друга) А.Н.Колмогоровым в 1933 году и В.И.Арнольдом в 2003 году.

В статье показано, что эти параметры, кажущиеся независимыми характеристиками поля случайных точек, становятся функционально зависимыми, когда их значения усреднены по малым флуктуациям точек поля.

Ключевые слова.

Комбинаторика — случайные распределения — притяжение — отталкивание — арифметические датчики случайных чисел — распределение Колмогорова — релятивистский метод теории вероятностей — конечная дискретная окружность.