

ИССЛЕДОВАНИЕ ЭФФЕКТА СУПЕР-ПОПУЛЯРНОСТИ В МУЗЫКАЛЬНОЙ ИНДУСТРИИ НА ОСНОВЕ ДАННЫХ SPOTIFY

Д. Д. Шабанова

*Саратовский национальный исследовательский
государственный университет им. Н.Г. Чернышевского, Россия*
E-mail: shabanovadaria04@gmail.com

Статья посвящена количественной оценке феномена сверхконцентрации популярности в музыкальной индустрии на основе данных платформы Spotify. Цель работы состоит в анализе асимметрии распределения популярности артистов с использованием закона Ципфа, а также методов статистического анализа данных с тяжелыми хвостами. Полученные результаты подтверждают степенную закономерность в хвосте распределения и позволяют выдвинуть гипотезу о его смешанном характере. Это служит основой для применения новых подходов к анализу данных медиаиндустрии.

RESEARCH ON THE SUPER-POPULARITY EFFECT IN THE MUSIC INDUSTRY USING SPOTIFY DATA

D. D. Shabanova

The article presents a quantitative assessment of the phenomenon of hyper-concentration of popularity in the music industry based on data from the Spotify platform. The objective of the work is to analyze the asymmetry in the distribution of artists' popularity using Zipf's law, as well as methods for the statistical analysis of heavy-tailed data. The obtained results confirm a power-law pattern in the tail of the distribution and allow for advancing a hypothesis regarding its mixed nature. This serves as a basis for applying new approaches to the analysis of media industry data.

Исследование закономерностей распределения успеха в музыкальной индустрии представляет научный и практический интерес. «Феномен суперзвезды» впервые описан в работе [1] американским экономистом Шервином Розеном. Он подразумевает, что подавляющая доля прослушиваний и доходов приходится на крайне малую группу людей, и является характерным признаком современных цифровых платформ. В данной работе выдвигается гипотеза о том, что распределение популярности артистов подчиняется степенному закону (закону Ципфа) – модели, описывающей системы с экстремальной асимметрией. Для проверки гипотезы используются методы статистического анализа данных с тяжелыми хвостами на основе данных, собранных с платформы Spotify.

Рассмотрим степенное распределение, задаваемое для непрерывной случайной величины X функцией плотности

$$f(x) = \frac{\alpha-1}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\alpha}, \text{ для } x \geq x_{min}, \alpha > 1,$$

где α – показатель степенного закона, а x_{min} – нижняя граница, начиная с которой закон выполняется.

Оценка параметра α степенного закона проводится методом максимального правдоподобия (ММП) и может быть получена в виде:

$$\hat{\alpha} = 1 + \frac{n}{\sum_{i=1}^n \ln(\frac{x}{x_{min}})}$$

Для визуализации данных с тяжелыми хвостами мы применяем логарифмический бининг – метод построения гистограммы, при котором ширина бинов (интервалов) увеличивается экспоненциально.

Проверка гипотезы о соответствии эмпирических данных степенному распределению в исследовании будет осуществляться с помощью критерия Колмогорова-Смирнова (KS). Статистика критерия имеет вид:

$$D = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|,$$

где $F_n(x)$ – эмпирическая, а $F(x)$ – теоретическая функции распределения. Нулевая гипотеза H_0 отвергается на уровне значимости α^* , если соответствующее p – value = $P(D \geq D_{\text{набл}} | H_0) < \alpha^*$, где $D_{\text{набл}}$ – наблюдаемое значение статистики Колмогорова-Смирнова. Критическое значение x_{min} подбирается автоматически как значение, максимизирующее p -value статистики KS [2].

Для анализа концентрации популярности в ранжированных выборках применяется закон Ципфа – частный случай степенного закона для рангов. Закон Ципфа утверждает, что частота появления элемента x , занимающего r -е место в ранжированном списке, обратно пропорциональна его рангу r

$$x(r) \sim \frac{1}{r^\beta},$$

где $x(r)$ – частота или популярность элемента с рангом r , β – показатель Ципфа, что в логарифмической форме имеет вид

$$\log x(r) = -\beta \log r + \text{const}.$$

Для данных, подчиняющихся степенному закону с показателем степени α , показатель закона Ципфа приближенно связан с ним соотношением [3]

$$\beta \approx \frac{1}{\alpha - 1}.$$

Эмпирический анализ в исследовании был выполнен на основе открытых данных платформы Spotify. Рассматриваемый датасет представляет собой рейтинг 3000 наиболее прослушиваемых артистов за всю историю платформы, включающий суммарное число прослушиваний для каждого исполнителя. Значения варьируются от 1.168 млрд до 119.811 млрд с медианой – 2.282 млрд и средним 4.434 млрд, что характеризует сильную правостороннюю асимметрию, где большинство артистов имеют относительно небольшое количество прослушиваний, в то время как небольшая группа демонстрирует экстремально высокие показатели. На рис. 1 представлена гистограмма, построенная с использованием логарифмического бининга. Такой метод позволяет визуализировать «тяжелый хвост» распределения.

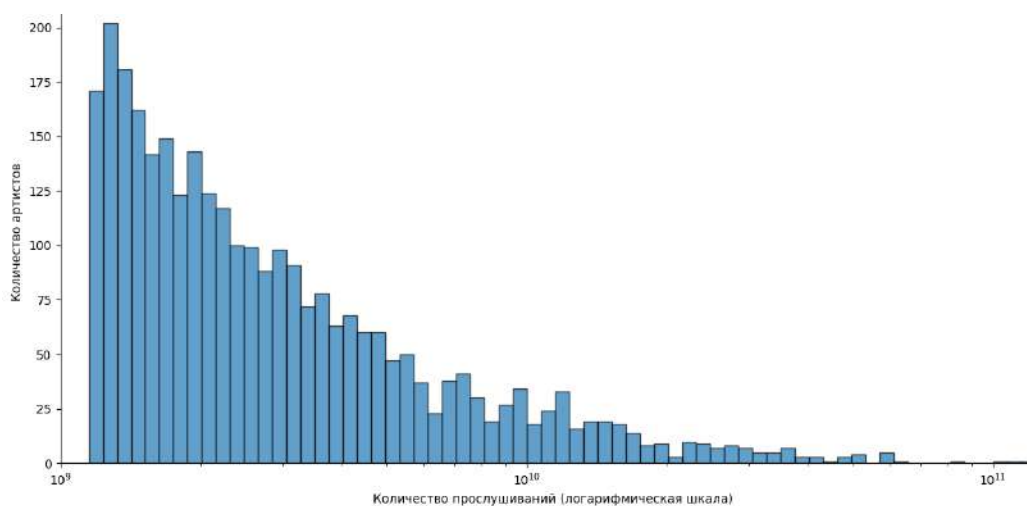


Рис. 1. Гистограмма распределения артистов по суммарному числу прослушиваний

Для проверки выдвинутой гипотезы был применен ММП с автоподбором нижней границы x_{min} , начиная с которой степенной закон выполняется. Описанный выше алгоритм определил оптимальный порог $x_{min} = 9.2$ млрд. прослушиваний, что составило 10% от общего объема выборки. Для исполнителей, превышающих этот порог, оценка параметра степенного закона $\hat{\alpha} = 2.69$. Статистика KS составила $D = 0.061$, а соответствующее значение p – $value = 0.201$. Таким образом, при $\alpha^* = 0.05$ гипотеза о том, что распределение популярности для топовых исполнителей следует степенному распределению, не отвергается.

Отсюда следует, что функция плотности данного распределения имеет вид

$$f(x) = \frac{1.69}{9.2 \cdot 10^9} \left(\frac{x}{9.2 \cdot 10^9} \right)^{-2.69}.$$

Полученная оценка $\hat{\alpha} = 2.69$ ($2 < \alpha \leq 3$) соответствует распределению с конечным математическим ожиданием, но бесконечной дисперсией. Это говорит о том, что количественная оценка разброса в популярности настолько велика, что ее невозможно описать с помощью классических статистических инструментов, а супер-популярность единичных исполнителей вносит доминирующий вклад, фактически формируя основную долю общего объема прослушиваний.

Для того же набора данных был проведен анализ, основанный на ранговом распределении [4]. На рис. 2 представлена зависимость логарифма числа прослушиваний $\log x(r)$ от логарифма ранга $\log r$. В результате аппроксимации данных линейной моделью была получена оценка показателя Ципфа $\beta = 1.24$ (коэффициент детерминации $R^2 = 0.982$).

На основе вычисленных коэффициентов уравнение регрессии в логарифмических координатах принимает вид

$$\log x(r) = -1.24 \cdot \log r + 34.07.$$

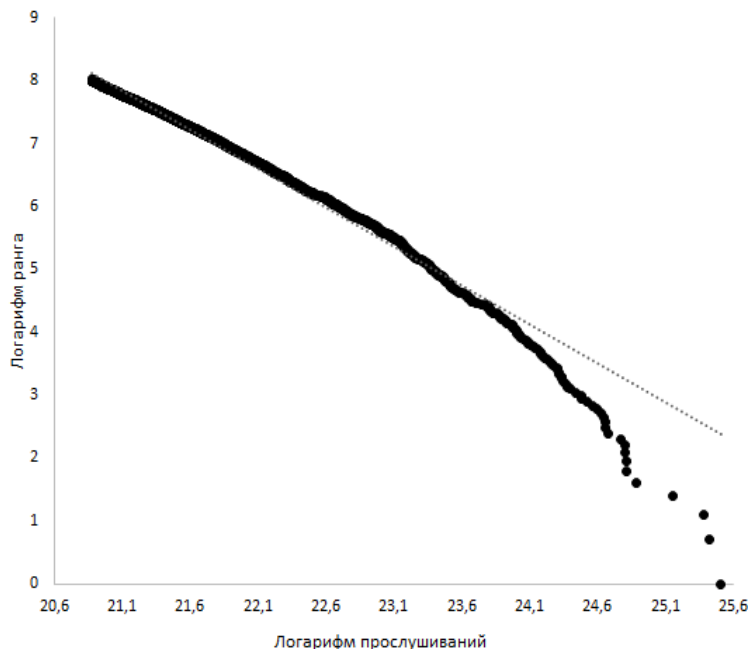


Рис. 2. Зависимость «ранг – размер» по показателю числа прослушиваний в логарифмических координатах

Полученная оценка $\beta = 1.24$ превышает каноническое значение $\beta_z = 1$. Это указывает на то, что популярность убывает с ростом ранга быстрее, чем в классической формулировке закона, что служит количественным подтверждением наличия «эффекта суперзвезды».

Подставив вычисленное значение $\alpha = 2.69$ в соотношение $\beta \approx 1/(\alpha - 1)$, получим теоретическое ожидание

$$\beta_{\text{теор}} \approx \frac{1}{2.69 - 1} \approx 0.59.$$

Видно, что теоретический показатель отличается от эмпирического значения β более чем в два раза. Такое расхождение является статистически значимым и указывает на сложную, неоднородную структуру рынка. Фактический параметр Ципфа $\beta > 1$ свидетельствует о том, что конкуренция и неравенство среди абсолютных лидеров музыкальной индустрии еще более экстремальны, чем это следует из общей модели степенного распределения, демонстрируя эффект сверхконцентрации в верхней части ранжированного списка.

Таким образом, распределение популярности артистов не может быть адекватно описано однородным степенным законом. Более точную модель может представлять собой смешанное распределение, в котором: «тело» соответствует экспоненциальному закону, а «хвост» – степенному.

В результате работы была количественно подтверждена гипотеза о соответствии распределения числа прослушиваний степенному закону в хвосте ($\alpha = 2.69$). Оценка показателя закона Ципфа ($\beta = 1.24$) так же подтвердила факт неравномерности концентрации успеха среди исполнителей. Расхождение полученных значений стало основанием для выдвижения гипотезы о необходимости

использования смешанной модели при описании структур подобного типа. Примером может служить двухкомпонентная модель, которая предполагает четкое разделение распределения на «тело» и «хвост»

$$f(x) = \begin{cases} A\lambda e^{-\lambda x}, & \text{если } x < x_c, \\ Bx^{-\alpha}, & \text{если } x \geq x_c, \end{cases}$$

где x_c — точка соединения, которая может соответствовать нижней границе степенного закона $x_{min} = 9.2$ млрд, а A и B — нормировочные константы.

Полученные результаты позволяют сформулировать основной вывод: для управления рисками в условиях сверхконцентрированного рынка необходимо переходить к сегментированным моделям, учитывающим разные режимы распределения успеха. Такой подход способен повысить адекватность принятия решений в медиандустрии.

СПИСОК ЛИТЕРАТУРЫ

1. *Rosen S.* The Economics of Superstars // *The American Economic Review*. 1981. Vol. 71, No. 5. P. 845–858.
2. *Clauset A., Shalizi C. R., Newman M. E. J.* Power-Law Distributions in Empirical Data // *SIAM Review*. 2009. Vol. 51, No. 4. P. 661–703.
3. *Newman M. E. J.* Power laws, Pareto Distributions and Zipf's Law // *Contemporary Physics*. 2005. Vol. 46, No. 5. P. 323–351.
4. *Zipf G. K.* Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology // Cambridge, MA : Addison-Wesley Press, 1949. 573 p.