

О законе Лотки и «среднем ученом»

Игорь Пшеничников

ЛРЯФ ОЭФ
ИЯИ РАН



Ученый совет ИЯИ РАН
07.04.2016

**Взаимоотношения между учёными,
преподавателями, руководителями институтов и
МОН, ФАНО, научными фондами**



Примитивизация информации при её движении наверх к начальству: от содержания и сути достижений к средним значениям и интегралам

- Научные публикации, содержательные описания достижений и результатов заменяются сводными таблицами количества статей, средними значениями, интегральными величинами.
 - Не понимая научных результатов по существу управленцы стремятся описать их простой количественной мерой для удобства подсчета и доклада начальству.
 - Мало надежды на то, что МОН и ФАНО откажутся от использования формальных количественных показателей для оценки подведомственных им учреждений.
-

Средние показатели научной результативности широко используются в документах ФАНО и МОН

- Пресс-релиз Совета по науке при Министерстве образования и науки РФ, 05.02.2015:
«Совет отмечает необходимость сохранения в п.13 положения об аттестации соотношения показателей аттестуемого работника либо со средними показателями по референтной группе организаций, либо со средними показателями для данного института.»
 - Газета «Поиск» № 40(2015) 05.10.2015 об итогах НКС ФАНО:
«Критерии определения лидеров в сети подведомственных ФАНО научных организаций: организация может быть признана лидером только в том случае, если ее показатели значительно превышают средние по референтной группе.»
-

Смысл и адекватность использования средних значений (дисперсии, медианы) можно понять и оценить только изучив само распределение

- Как показатели результативности (успешности) распределены среди членов научного сообщества? Можно ли определить лидеров и отстающих путем сравнения их показателей со средними по институту?
 - Правильно ли оценивать институты путем сравнения средних значений, вычисленных по каждому институту в целом?
-

Публикации: Закон Лотки, $x_k \sim 1/k^2$

- В 1926 году Альфред Лотка исследовал распределение продуктивности физиков и химиков в виде количества опубликованных ими статей за 10 лет: A.J. Lotka, The frequency distribution of scientific productivity, J. Wash. Acad. Sci. 16 (1926) 317
- Сформулированный им эмпирический закон гласит, что число x_k учёных, каждый из которых написал k статей, можно аппроксимировать зависимостью $x_1 * k^{-\alpha}$ со значениями α , близкими к 2.
- Следуя этой зависимости можно ожидать, что если в научном коллективе есть 100 ученых, которые написали по одной статье за определенный промежуток времени, то в этом же коллективе будут примерно 25 других ученых, которые написали по две статьи каждый, будет около 11 учёных — с тремя статьями, шесть ученых — с четырьмя статьями и т. д. Наконец, можно ожидать, что будет только один автор с 10 публикациями.
- За минувшие 90 лет справедливость закона Лотки была в целом подтверждена многими десятками работ, где значения показателя α от 1.4 до 3.9 были использованы для описания аналогичных наборов данных.

Цитирования: степенная зависимость

PHYSICAL REVIEW E **68**, 026113 (2003)

Citation networks in high energy physics

S. Lehmann, B. Lautrup, and A. D. Jackson

The Niels Bohr Institute, Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark

(Received 16 December 2002; revised manuscript received 20 March 2003; published 15 August 2003)

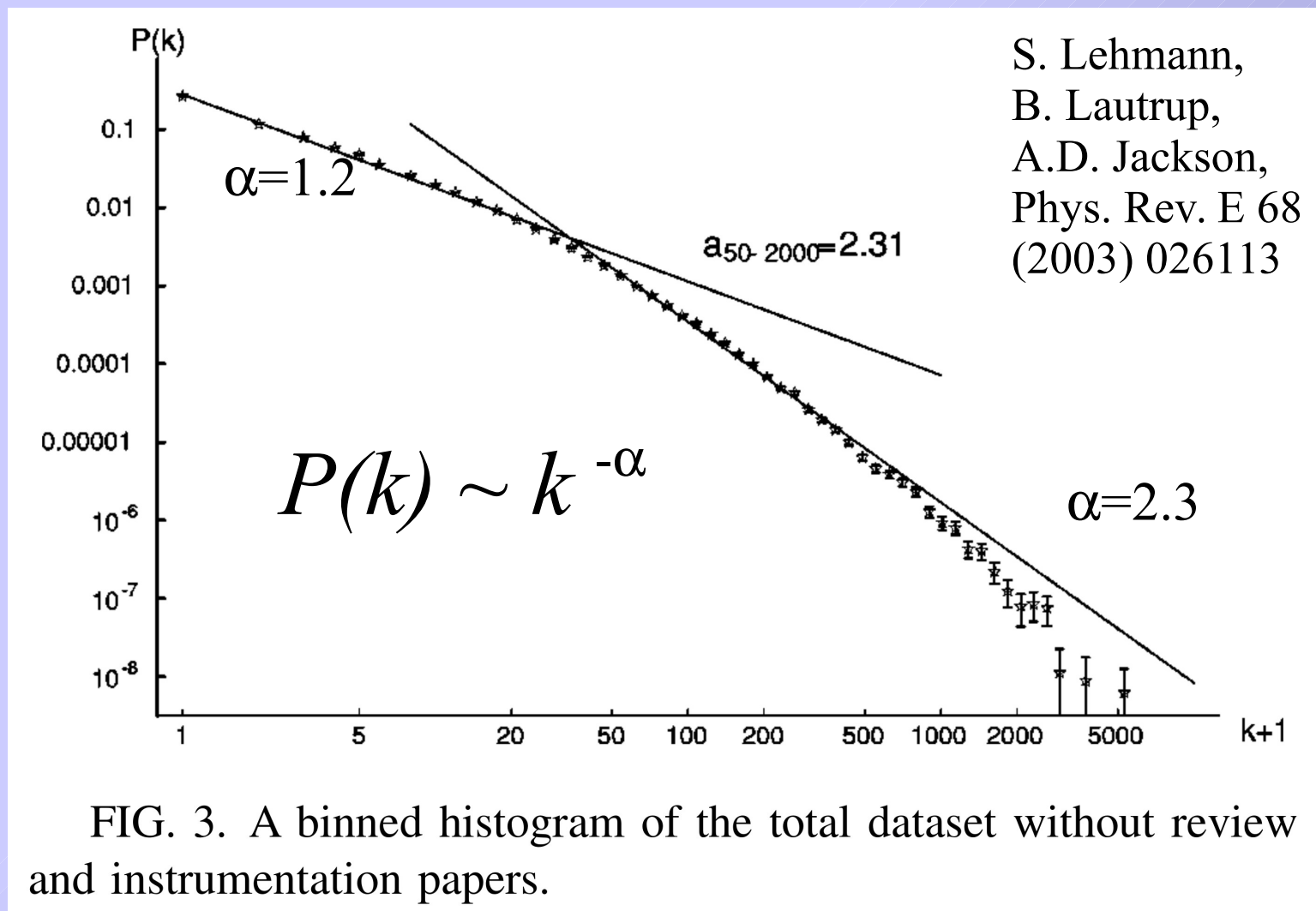
The citation network constituted by the SPIRES database is investigated empirically. The probability that a given paper in the SPIRES database has k citations is well described by simple power laws, $P(k) \propto k^{-\alpha}$, with $\alpha \approx 1.2$ for k less than 50 citations and $\alpha \approx 2.3$ for 50 or more citations. A consideration of citation distribution by subfield shows that the citation patterns of high energy physics form a remarkably homogeneous network. Further, we utilize the knowledge of the citation distributions to demonstrate the extreme improbability that the citation records of selected individuals and institutions have been obtained by a random draw on the resulting distribution.

DOI: 10.1103/PhysRevE.68.026113

PACS number(s): 89.65.-s, 89.75.-k

Частотное распределение числа цитирований k работ из международной базы публикаций SPIRES (ныне INSPIRE-HEP, <http://inspirehep.net>, физика частиц высоких энергий, астрофизика, космология) удивительно хорошо описывается степенными зависимостями $P(k) \sim k^{-\alpha}$

Распределение вероятности $P(k)$ для работы из SPIRES (кроме обзоров) быть процитированной k раз



4% работ получили 50% всех цитирований, а менее удачливая половина всех работ получила только 2% всех цитирований.
Большая разница между средними и медианными значениями.

Интересно, что совпадают распределения $P(k)$ для различных категорий: «теория», «феноменология», «эксперимент»

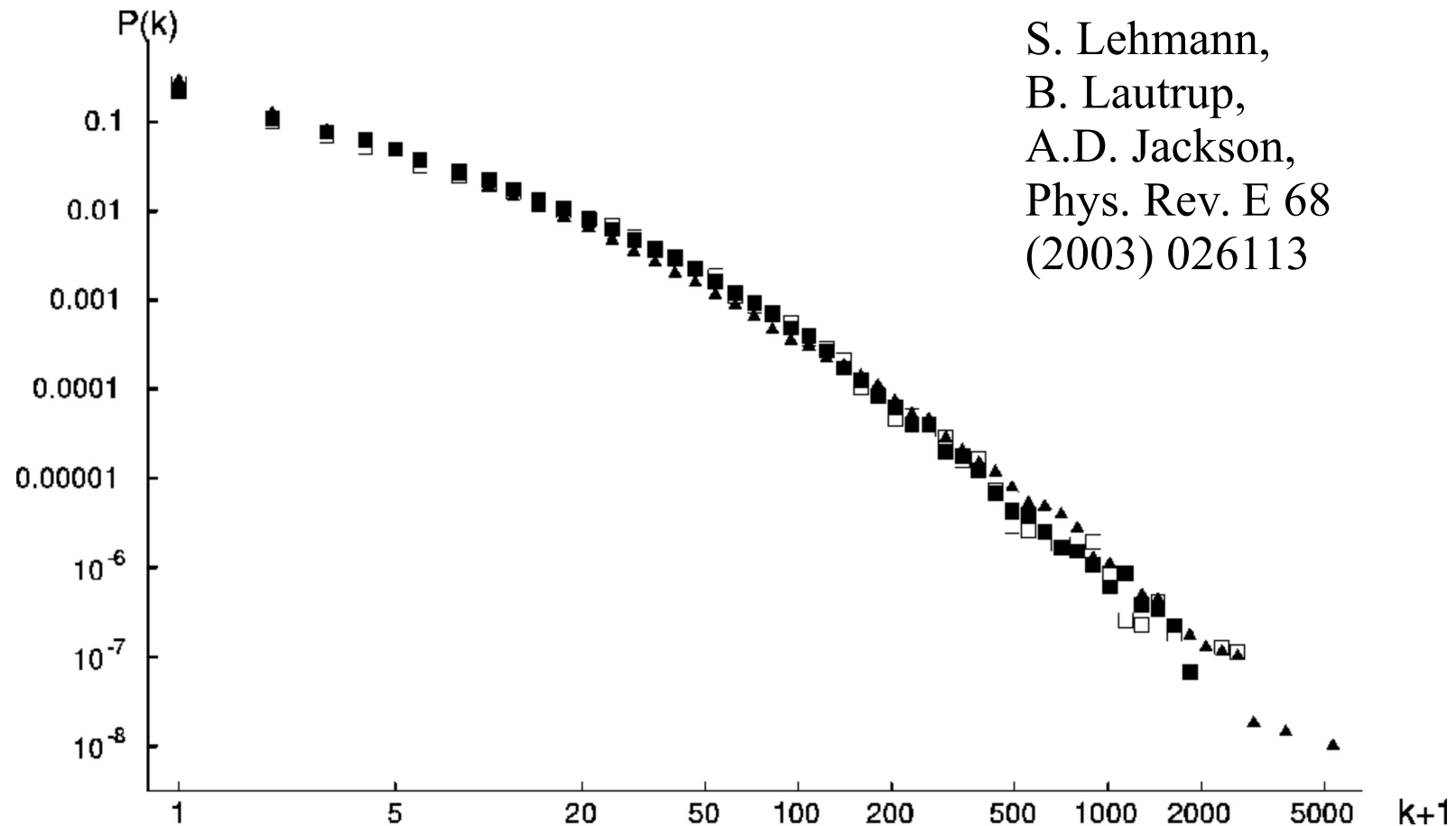
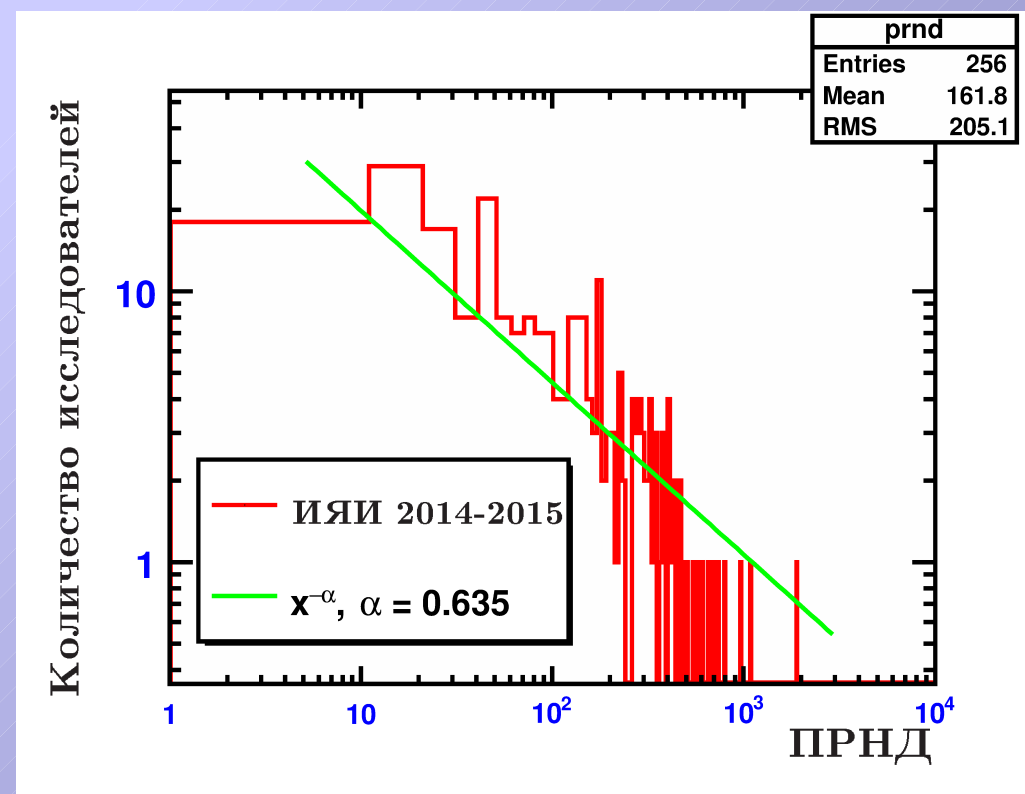
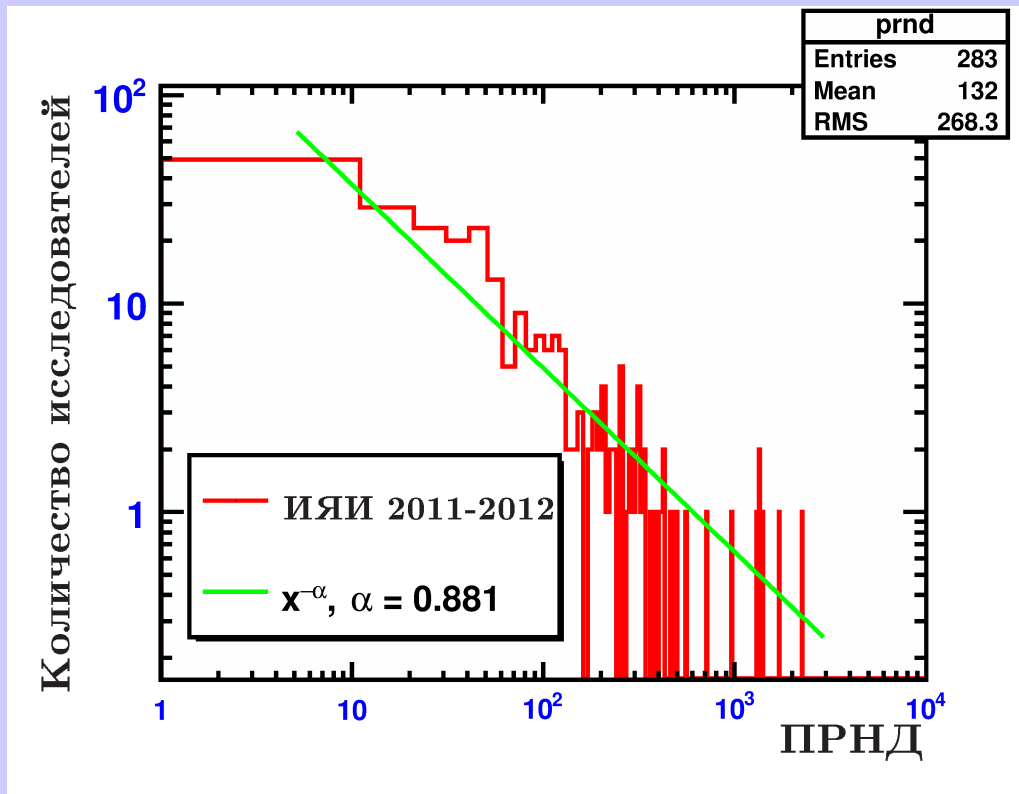


FIG. 2. Degree distributions for the categories theory ($\blacktriangle$), phenomenology ($\blacksquare$), and experiment ($\square$).

ПРНД сотрудников ИЯИ РАН в 2013 и 2016 годах

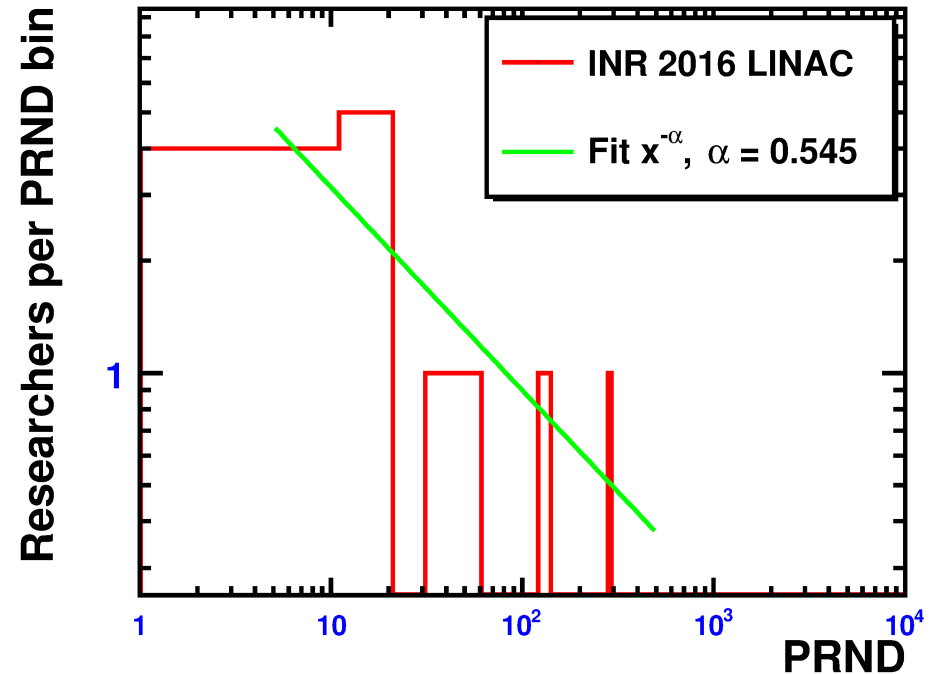
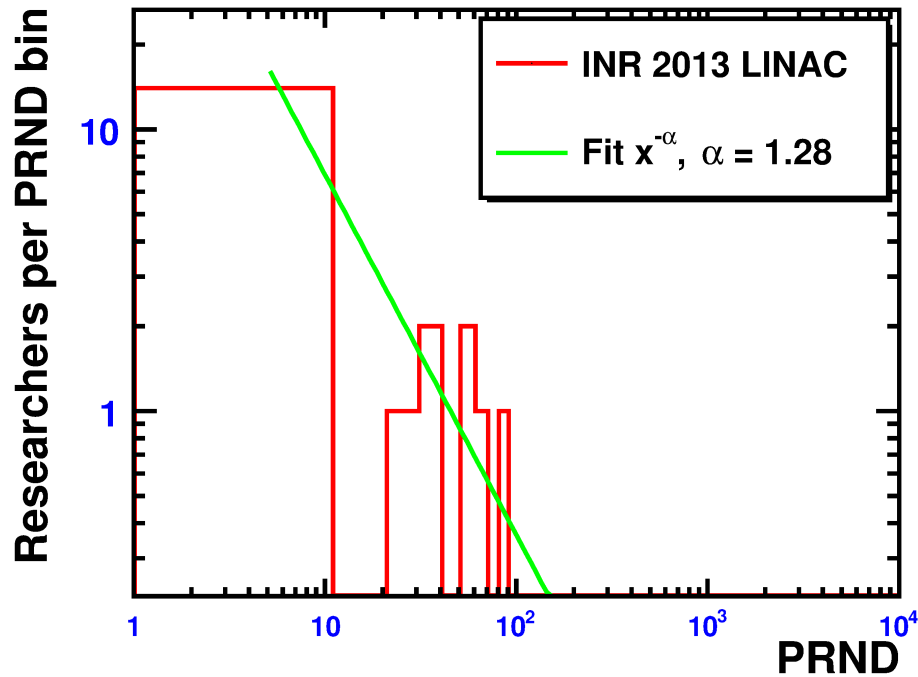


ПРНД — обобщённый показатель результативности, произведение количества публикаций и IF журналов (качества).

В целом неплохо описывается зависимостью $x^{-\alpha}$

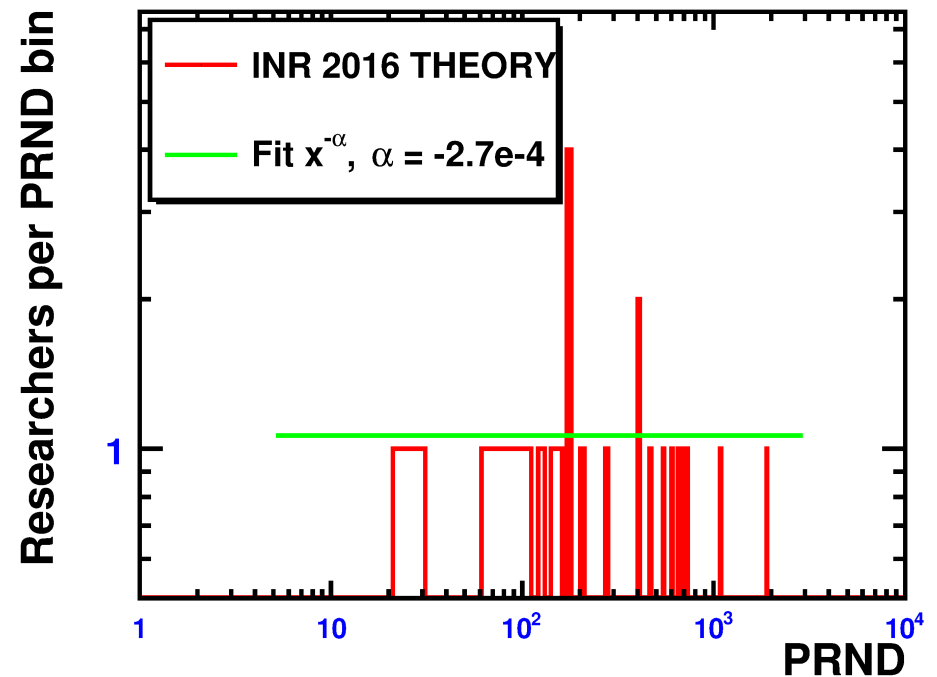
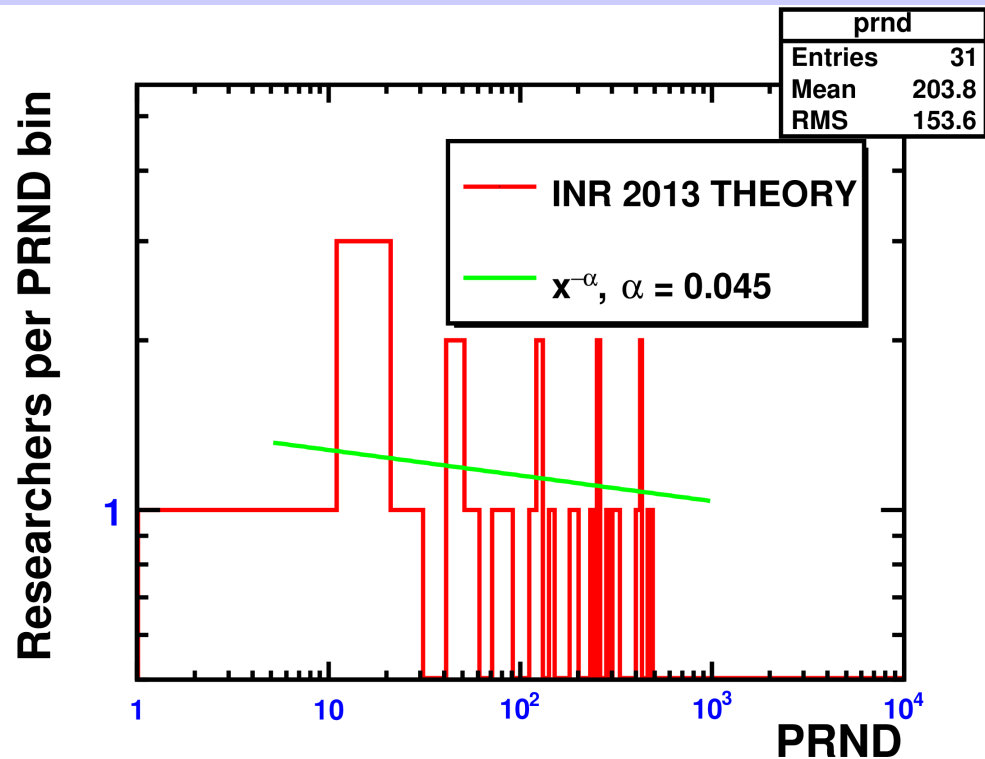
Разница в величине α может быть связана с изменениями правил ПРНД в 2016 г.

ИЯИ Отдел ускорителей 2013 и 2016



Малый размер выборки, но какие-то следы степенной зависимости при желании можно увидеть и здесь.

ИЯИ Теоретдел 2013 и 2016



Малый размер выборки. Аппроксимация степенной зависимостью теряет смысл, происходит кластеризация вокруг некоторых средних значений.

Заметное число работ публикуется вместе с соавторами из того же отдела — высокая степень корреляции ПРНД.

Разнообразные причины вариаций формальной результативности

- Количество публикаций тесно связано с трудоемкостью и ресурсоёмкостью исследований по данной теме или научному направлению, количеством соавторов
- Большие экспериментальные установки создаются десятилетиями, в течении которых выходят только немногочисленные публикации по методике. Только после начала работы идет поток публикаций
- Цитируемость зависит от «населённости» конкретной области исследований. Статистика EXFOR: измерения с U, C vs Po, Am, Bk

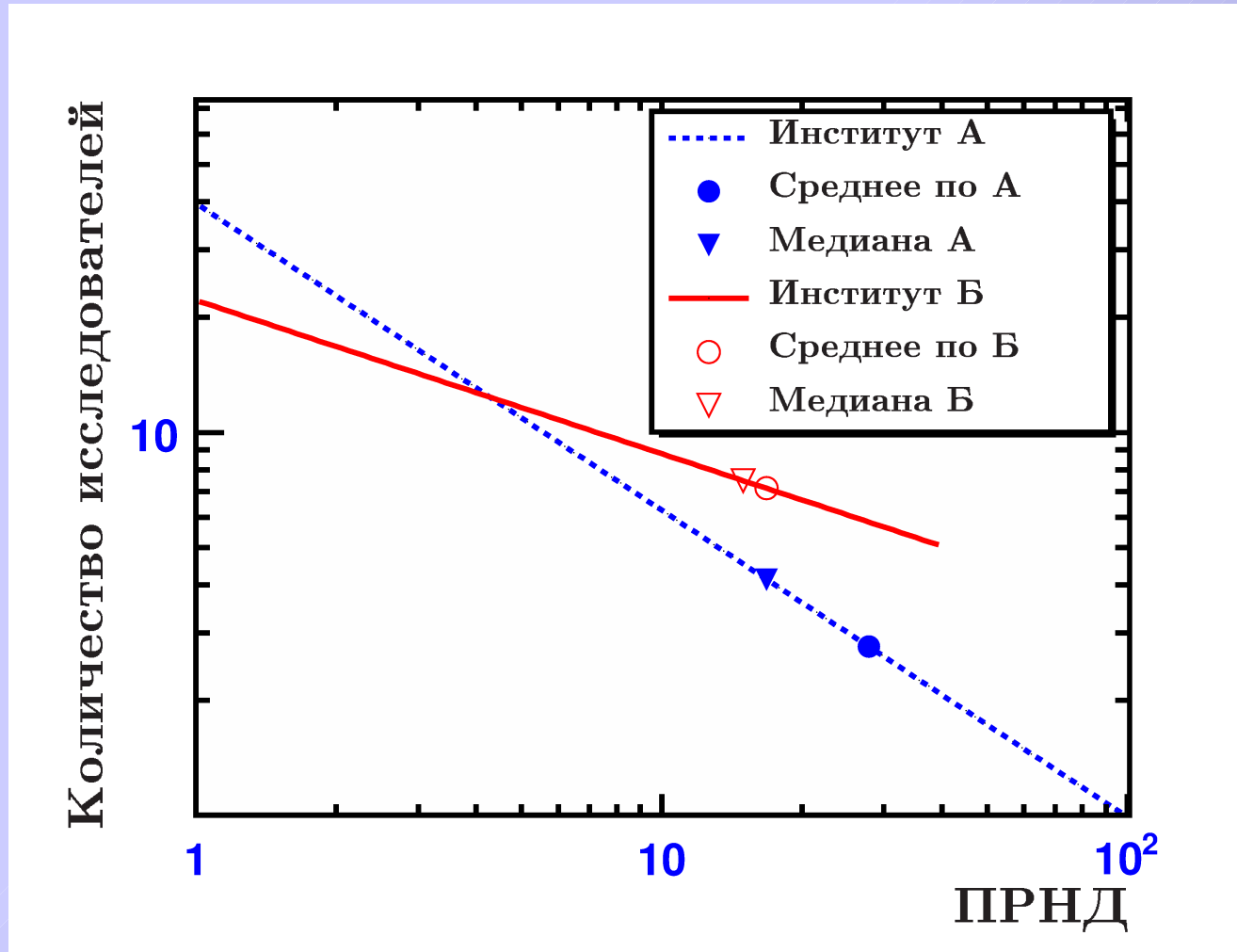
EXFOR Targets (el.)

#	Target	Counts	Percent
1	U	2496	11.8
2	C	1948	9.22
3	H	1579	7.48
4	Al	1344	6.36
5	Ni	1297	6.14
6	Fe	1271	6.02
7	Pb	1246	5.9
8	Cu	1056	5
9	Li	1000	4.73
10	Pu	939	4.44
11	Co	922	4.36

92	Pm	27	0.12	32
93	Fm	26	0.12	79
94	Bk	24	0.11	77
95	Es	17	0.08	76
96	NN	16	0.075	54
97	Ac	12	0.056	4
98	No	8	0.037	26
99	Po	5	0.023	26
100	Md	3	0.014	22
101	Rf	3	0.014	21
102	Db	2	0.0094	1
103	ELEM/MASS	2	0.0094	22
104	Rn	2	0.0094	
105	Lr	1	0.0047	

Total: 105

Сравнение институтов А и Б



- Предположим, что в институтах А и Б работает по 300 сотрудников, а распределения ПРНД следуют зависимости $x^{-\alpha}$ с показателями $\alpha = 0.8$ и 0.4 , соответственно.
- Допустим, что диапазоны их ПРНД в них составляют, соответственно, от 1 до 100 и от 1 до 40.

Сравнение институтов А и Б: выводы

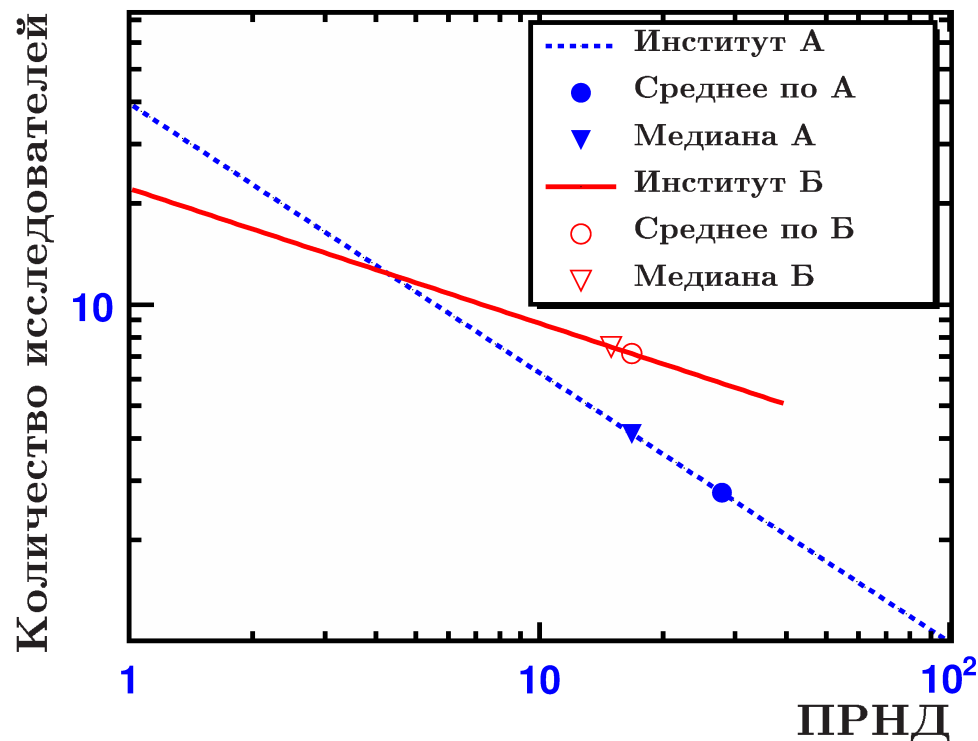
Средние значения ПРНД А: 27.8
Б: 16.8

поэтому следуя ФАНО и МОН,
Институт А — лидирует с заметным
преимуществом.

Однако средние ПРНД имеют
только несколько сотрудников и
их значения не характеризуют
весь коллектив в целом!

Сотрудников с ПРНД > 4
в институте А - 237, а в
более «слабом» институте Б — 252!

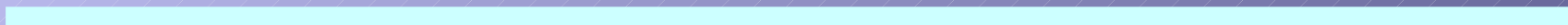
Аналогично, сравнение медианных
значений, (они близки у А и Б!) не
позволяет сделать однозначных
выводов.



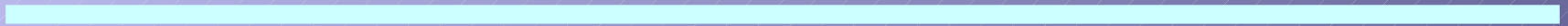
Вывод: сравнение со средними или медианными значениями при оценке результативности сотрудников и институтов не имеет смысла

- Степенные распределения характерны для разных областей знания и разных стран.
 - Следует правильно интерпретировать различия формальных показателей результативности, **которые варьируются в широких пределах в зависимости от характера научной деятельности внутри референтной группы институтов и даже внутри одного института.**
 - Для оценки институтов ФАНО следует обратить внимание на конкретные результаты работающих в них научных групп, лабораторий и отдельных учёных
 - **Недопустимо определять успешно работающие и отстающие институты или лаборатории путем сравнения средних или медианных показателей.**
-

Спасибо за внимание!



Backup slides



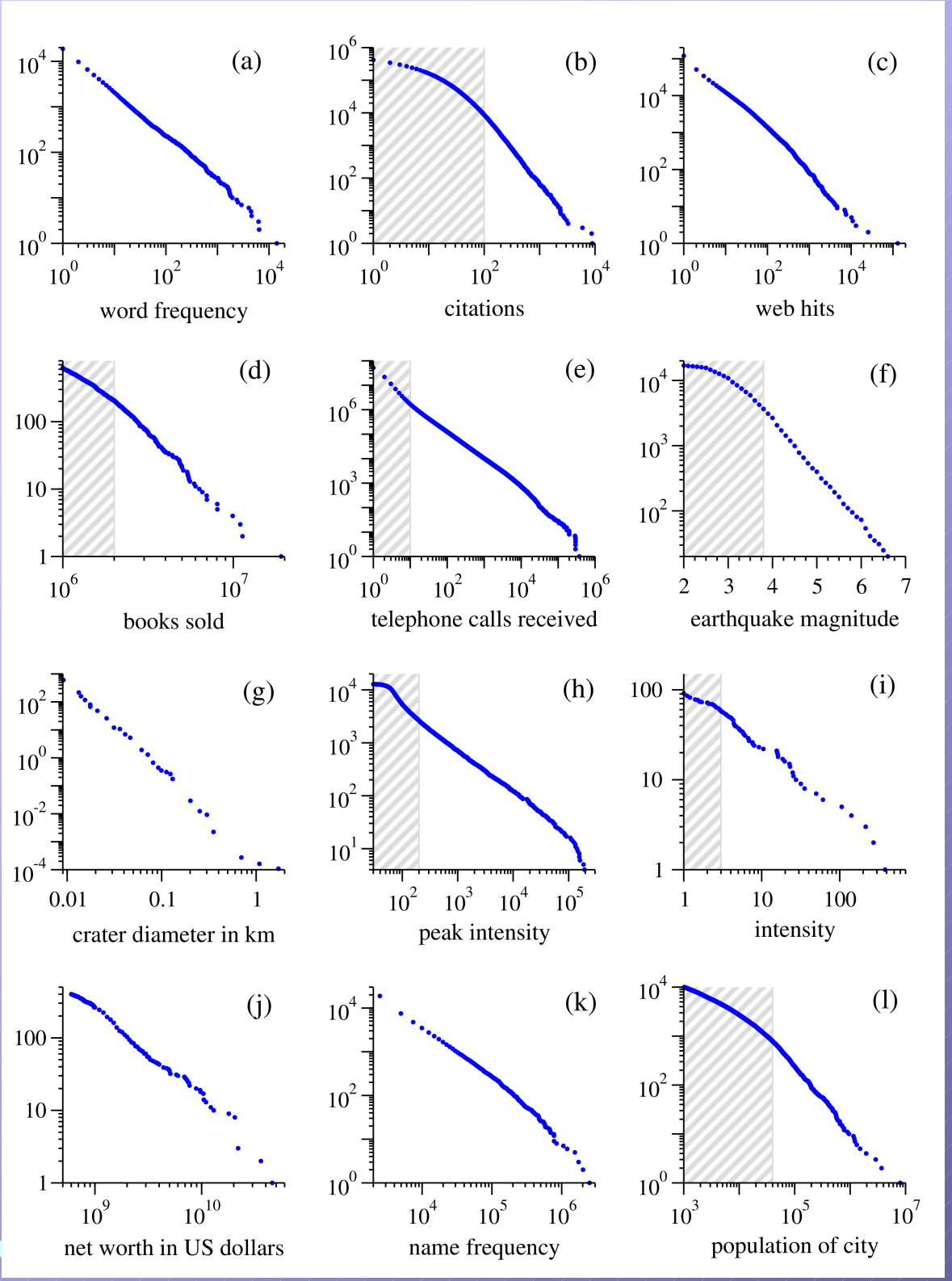
Степенные зависимости, распределение Паретто, закон Ципфа

quantity	minimum $x_{\min}$	exponent α
(a) frequency of use of words	1	2.20(1)
(b) number of citations to papers	100	3.04(2)
(c) number of hits on web sites	1	2.40(1)
(d) copies of books sold in the US	2 000 000	3.51(16)
(e) telephone calls received	10	2.22(1)
(f) magnitude of earthquakes	3.8	3.04(4)
(g) diameter of moon craters	0.01	3.14(5)
(h) intensity of solar flares	200	1.83(2)
(i) intensity of wars	3	1.80(9)
(j) net worth of Americans	\$600m	2.09(4)
(k) frequency of family names	10 000	1.94(1)
(l) population of US cities	40 000	2.30(5)

TABLE I Parameters for the distributions shown in Fig. 4. The labels on the left refer to the panels in the figure. Exponent values were calculated using the maximum likelihood method of Eq. (5) and Appendix B except for the moon craters (g), for which only cumulative data were available. For this case the exponent quoted is from a simple least-squares fit and should be treated with caution. Numbers in parentheses give the standard error on the trailing figures.

M.E.J. Newman

<http://arxiv.org/abs/cond-mat/0412004v3>



Принципиальные отличия от распределений с характерным средним значением

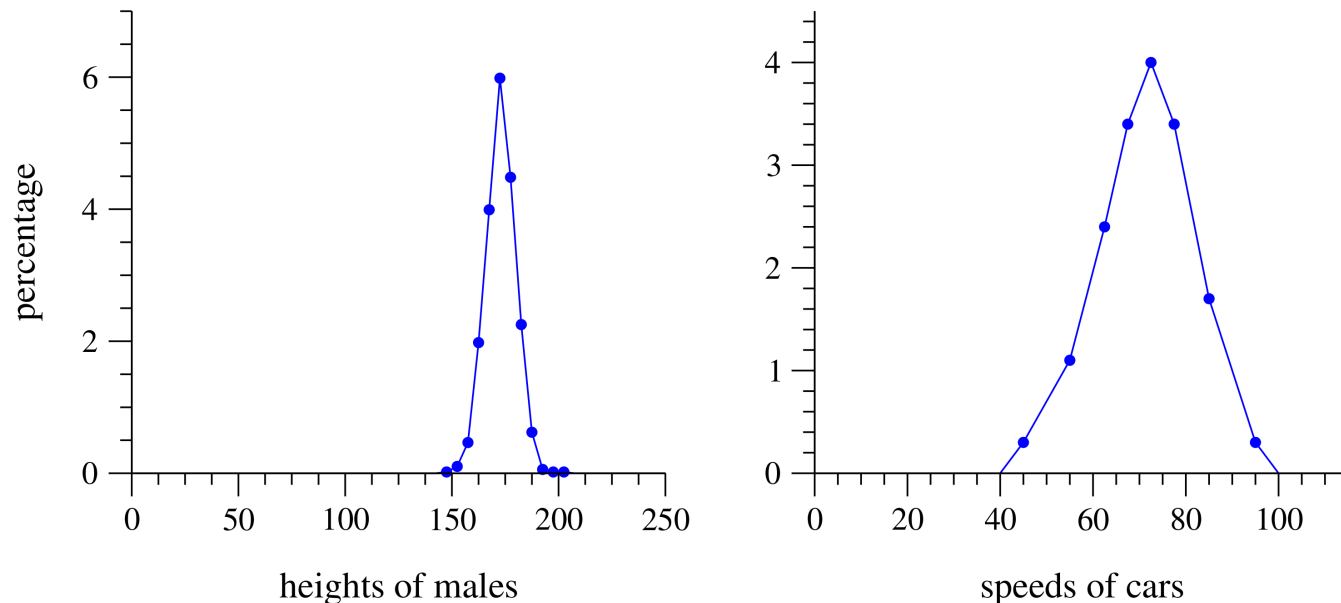


FIG. 1 Left: histogram of heights in centimetres of American males. Data from the National Health Examination Survey, 1959–1962 (US Department of Health and Human Services). Right: histogram of speeds in miles per hour of cars on UK motorways. Data from Transport Statistics 2003 (UK Department for Transport).

Распределение граждан мужского пола по росту

Распределение скоростей автомобилей, движущихся по автострадам

Свойства степенной зависимости: моменты распределения

- Среднее расходится при $\alpha < 2$, а среднеквадратичное расходится при $\alpha < 3$.

$$\begin{aligned}\langle x \rangle &= \int_{x_{\min}}^{\infty} x p(x) dx = C \int_{x_{\min}}^{\infty} x^{-\alpha+1} dx \\ &= \frac{C}{2-\alpha} \left[x^{-\alpha+2} \right]_{x_{\min}}^{\infty}.\end{aligned}$$

$$\langle x^2 \rangle = \frac{C}{3-\alpha} \left[x^{-\alpha+3} \right]_{x_{\min}}^{\infty}.$$

Среднее сильно зависит от точек на правом краю распределения — его определяют единичные лидеры!

- Поэтому нельзя ввести устойчивое понятие «среднего ученого» или его «средней» активности и типичного разброса показателей и определить диапазон значений, в которых заключено, скажем 80% «нормальных» или «стандартных» учёных.

Вероятность для индексируемой в SPIRES публикации (из 281 717 на 14.08.2002) получить определённое количество цитирований

TABLE I. The probability of a paper in the SPIRES database having k citations for $0 \leq k \leq 4$ as a function of subfield. The total number of papers in each subfield is 159 946 (theory), 68 549 (phenomenology), 28 527 (experiment), 19 637 (instrumentation), and 5058 (review papers). The “total” data entries are obtained directly from the subfield data. The total number of papers in the dataset is 281 717.

	$P(0)$	$P(1)$	$P(2)$	$P(3)$	$P(4)$
Theory	0.2884	0.1226	0.0815	0.0590	0.0472
Phenomenology	0.2150	0.1103	0.0762	0.0618	0.0488
Experiment	0.2677	0.1023	0.0704	0.0518	0.0441
Instrumentation	0.6169	0.1206	0.0622	0.0385	0.0267
Review articles	0.2167	0.1038	0.0670	0.0496	0.0403
Total	0.2901	0.1171	0.0775	0.0574	0.0458

TABLE II. The search option “citation summary” at the SPIRES website returns the number of papers for a given author in the categories in this table. The probabilities of getting citations in these are intervals are listed in the third column.

Paper category	Citations	Probability
Unknown papers	0	0.267
Less known papers	1–9	0.444
Known papers	10–49	0.224
Well-known papers	50–99	0.0380
Famous papers	100–499	0.0250
Renowned papers	500+	0.00184

Оказывается, что почти каждая третья публикация никогда не цитируется, даже без поправки на самоцитирования. Доли таких «мёртвых» работ в категориях «теория» и «эксперимент» почти совпадают.

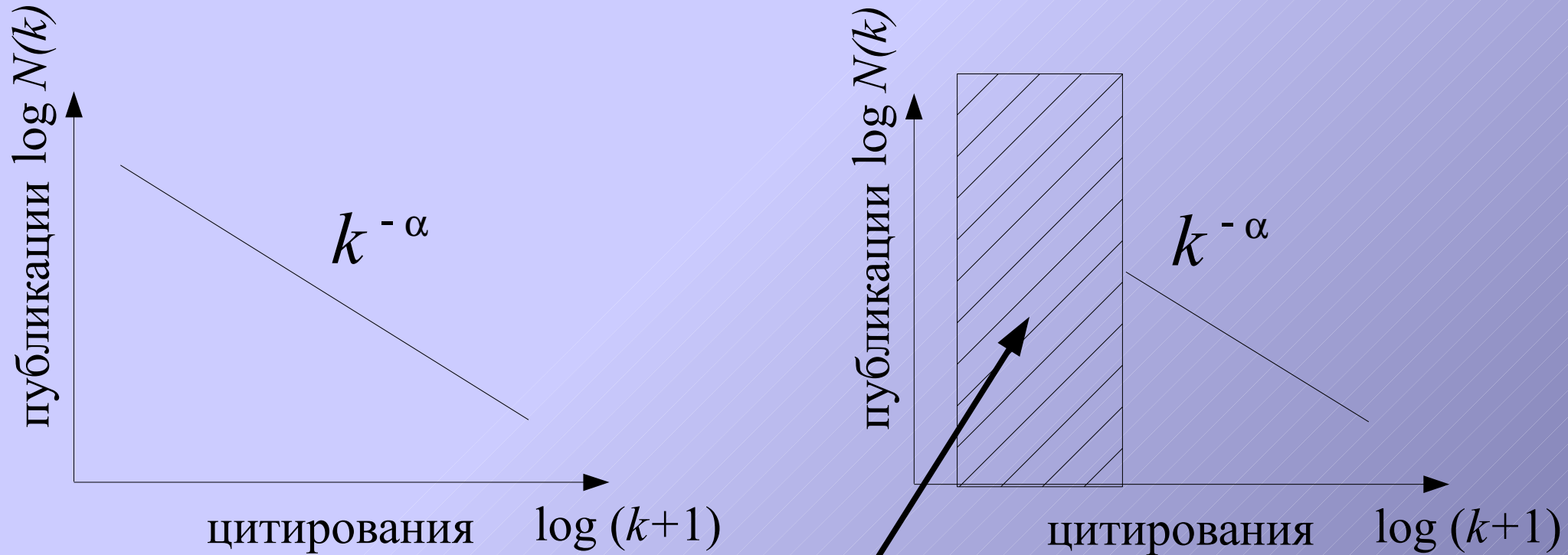
74% работ имеют 10 или менее цитирований

При этом только 6.2% статей цитируются 50 раз или более

И только 131 работа (~0.05%) цитируется 1000 или более раз

S. Lehmann, B. Lautrup, A.D. Jackson, Phys. Rev. E 68 (2003) 026113

«Реформа науки» с целью повышения формальной результативности

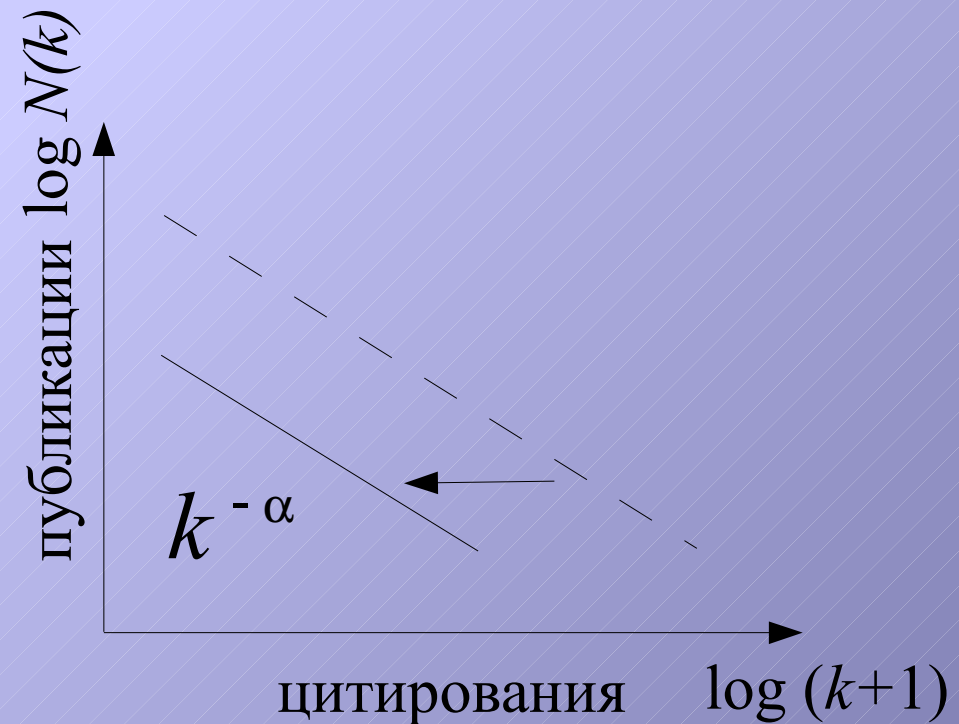
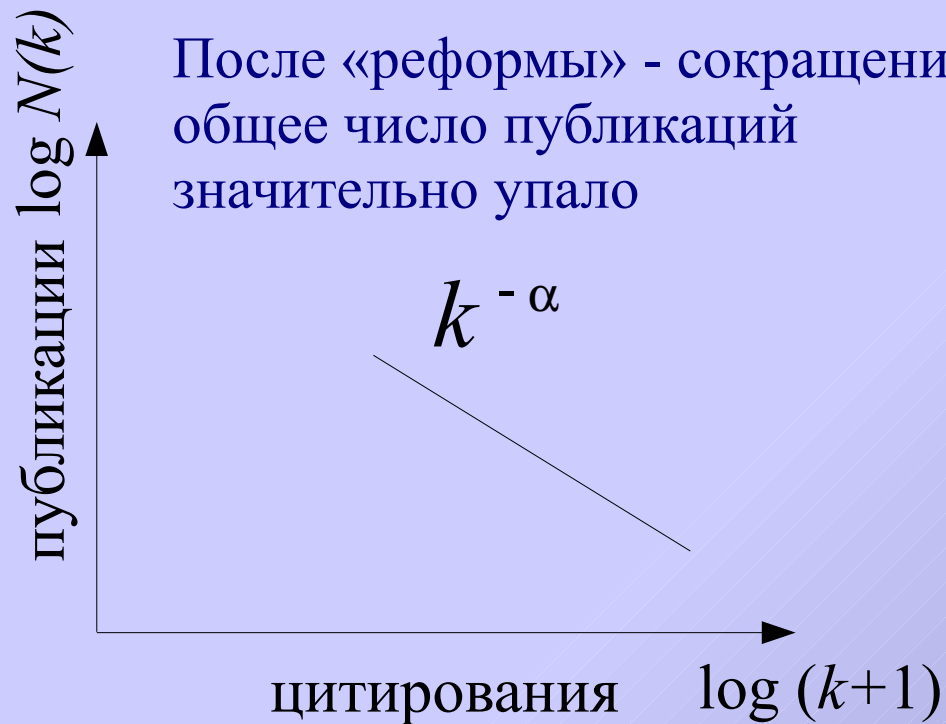


Важно: почти у каждого исследователя есть (относительно) малоизвестные работы

Проводим «реформу», увольняем исследователей, публикующих малоцитируемые работы

При этом сокращения проводим в разы, чтобы повлиять на среднее количество цитирований на статью

Плачевный результат «реформы»



Такая искусственно «улучшенная» зависимость будет сохраняться только короткое время после «реформы». Как было отмечено выше:

- (1) показатель степени - α един для областей «теория», «эксперимент», «феноменология»
- (2) всегда есть большое количество малоцитируемых работ

В результате последствий «реформы» вместе с общим количеством работ упадёт и количество высокоцитируемых работ!

Результат будет противоположен цели «реформы»!

О сокращениях в институтах ФАНО

- Небольшое (в сравнении лидерами в данной научной дисциплине) количество опубликованных научных работ и их цитирований не может служить основанием для увольнения научного работника
 - Если будет отрезана левая часть распределения, то будет разрушена база, из которой растут лидеры. В результате - через несколько лет - без подпитки снизу - осядет и лидерская часть.
 - Разумеется, научные коллективы должны избавляться от работников с нулевой результативностью, но каждый такой случай стоит рассматривать всесторонне и тщательно.
-