# Лекция 2. Идентификация параметров степенного распределения безмасштабных сетей.

## 1. Понятие безмасштабной сети. Свойства функции распределения степеней узлов.

### 1.1. Безмасштабная сеть.

Безмасштабная сеть или масштабно-инвариантная сеть ([англ.](https://dic.academic.ru/dic.nsf/ruwiki/6161) scale-free network) — [граф](https://dic.academic.ru/dic.nsf/ruwiki/8789), в котором степени вершин распределены по степенному закону, то есть доля вершин со степенью k примерно или асимптотически пропорциональна .

n – характеристическая степень, обычно между 2 и 3.

Безмасштабные сети самоподобны: в любом участке сети распределение степеней будет тем же.

Многие естественно возникающие сети – социальные, коммуникационные, графы цитирований, ссылок в Интернет и другие – хорошо моделируются безмасштабными графами.

Главная отличительная черта безмасштабных сетей –это существование узлов-концентраторов (хабов), степени которых очень велики по сравнению со степенями остальных узлов.

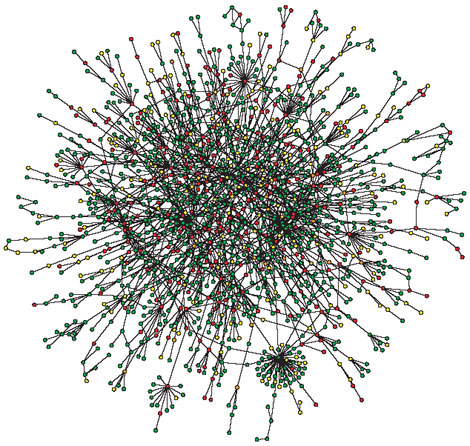


Рис. 1.1. Пример безмасштабной сети.

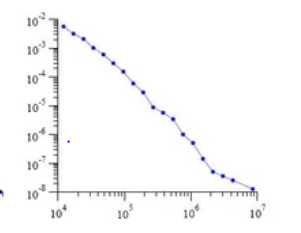
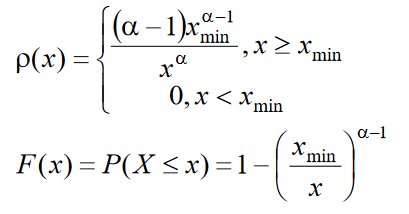


Рис. 1.2. Степенное распределение безмасштабной сети.

### 1.2. Свойства степенного распределения.

Интенсивное научное исследование проблематики обнаружения зависимостей степенного вида в выборочных данных различной природы традиционно ведет свой отсчет от работ социолога В. Парето. Именно ему принадлежит наблюдение, что распределение людей по доходам и/или накоплениям подчиняется степенному закону: доля тех, чьи доходы выше порога, описывается зависимостью , где α — показатель Парето. Схожие закономерности были найдены лингвистом Дж. Ципфом при изучении частоты встречаемости слов естественного языка. Подобные же степенные зависимости проявляются в целом ряде физических, биологических, социально-экономических систем. В общем случае распределение исследуемой переменной x будет иметь степенной характер(нами будет использоваться также термин «распределение Парето»), если ее плотность вероятности описывается как: (1)

К настоящему времени распределению Парето посвящена обширная литература, прежде всего, социально-экономического направления. Предлагаются различные динамические и статистические модели, объясняющие появление подобных распределений, часто именуемых распре-делениями с «тяжелым хвостом». Обсуждается более сложное распределение Леви, которое в частных случаях переходит в распределение Парето. Распределения с «тяжелыми хвостами» в реальных ситуациях играют существенную роль при анализе экономических показателей различного рода, при оценке вероятности катастроф и иных экстраординарных событий. Легко показать, что для больших x формула (1) дает результаты, на много порядков отличающиеся от аналогичной оценки, полученной из предположения о том, что x подчиняется нормальному, гауссовому распределению. Перечислим важнейшие свойства степенных распределений. При малых x и любых α выражение (1) неограниченно возрастает. Поэтому под «распределением Парето» обычно понимается распределение, для которого отсутствуют значения x меньше некоторого порога xmin. При этом условии выражения для плотности вероятности ρ(x) и функции распределения F(x) будут иметь вид:



На практике логично следовать соображению, что при x < xmin имеет место какое-либо другое распределение, а при x = xmin оно без разрыва переходит в распределение Парето. Уместно задаться вопросом, не будет ли более конструктивно подобрать такой класс распределений, которые могут быть пригодны для аппроксимации выборочных характеристик транзакционных данных во всем диапазоне возможных значений. Например, для гамма-распределения с помощью метода максимального правдоподобия могут быть рассчитаны коэффициент формы и коэффициент масштаба, удовлетворяющие наилучшему приближению выборочного распределения теоретическим. Не отрицая возможности такого решения, отметим тот факт, что Парето-распределение зависит от одного параметраα, величина которого может быть сравнительно просто и с достаточной достоверностью оценена из имеющихся в распоряжении исследователя массивов данных, причем и массивов весьма скромной размерности. Очевидно, что попытка оценить несколько параметров при недостатке данных может привести к тому, что корректно рассчитанные доверительные интервалы для оценок будут крайне широки.

### 1.3. Почему же сеть безмасштабная.

Для определения понятия и характеристик безразмерной сети необходимо ввести следующие понятия.

**Степенное распределение**

Вместо распределения числа связей по закону Пуассона, который имеет строгий максимум около среднего значения, для многих реальных сетей, например, таких как структур Интернета и его виртуального двойника WorldWideWeb, такого среднего значения не существует, а соответствующее вероятностное распределение подчиняется свойственному всем критическим состояниям степенному закону.

На рисунке 1.2 продемонстрировано: признак безмасштабных сетей – линейный график зависимости числа узлов от количества связей при построении в двойном логарифмическом масштабе.

**Нагрузка узла.**

Для многих безмасштабных сетей необходимо определить относительную важность входящих в нее узлов. Рассмотрим загруженность узла i в сети концентраторов, определяемую как суммарное число кратчайших путей между всеми остальными узлами, которые проходят через (узел) i. Эта величина также важна и при изучении транспортных потоков, обычно, называется нагрузкой (загруженностью) узла (или связи), поскольку характеризует долю проходящих через узел кратчайших путей и узлы с высоким значением B являются наиболее загруженными. В отличие от степени узла, понятие важности узла отражает топологию всей сети.

**Средняя длина пути.**

Средняя длина пути между узлами в безмасштабной сети увеличивается в среднем, как логарифм размера сети.

**Коэффициент кластеризации.**

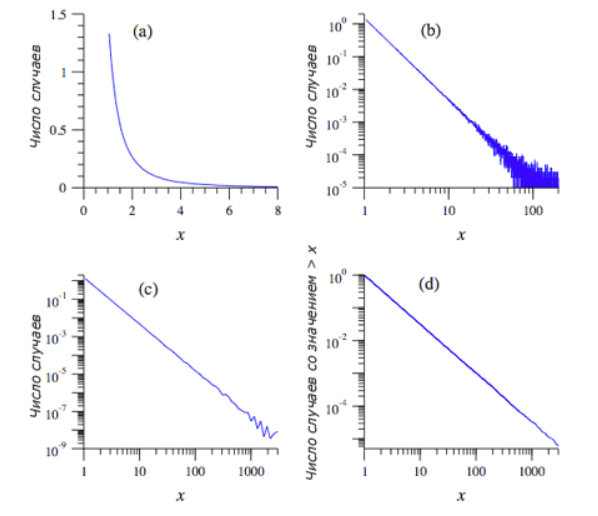
Коэффициенты кластеризации, полученные эмпирическим путем, в общем случае значительно выше для безмасштабных сетей, чем для случайных сетей. Коэффициент кластеризации зависит от размера сети согласно приближенному степенному закону. Это поведение отличается от поведения малых сетей, где кластеризация не зависит от размера сети.

Наличие большого числа интерпретаций модели безмасштабной сети (модель Боллобаша - Риордана, модель Барабаши – Альберт со степенным распределением, модель Пападопулуса и др.) говорит о том, что достаточно сложно построить математическую интерпретацию безмаштабной сети для практического исследования и построения модели. Каждая из предложенных моделей обладает своими характеристиками, особенностями и недостатками. Однако, все данные модели не являются полностью описанными математически, так, допустим, для каждой из моделей не посчитано математическое ожидание (среднее) количества ребер, связывающих вершины со степенями.

Таким образом, не смотря на ценность моделей безмасштабных сетей, их практическое применение является затруднительным (т.е. по данным моделям можно построить безмастшабную сеть, но практически сложно оценить её характеристики). Исходя из этого, необходимо предложить такую модель сети, которая бы не только практически соответствовала характеристикам сети, но и позволяла эффективно производить практические вычисления. Такой моделью можно считать иерархическую модель безмасштабной сети.

### 1.4. График функции распределения степеней узлов в логарифмических координатах.

Идентификация степенных законов в натуральных или искусственных системах не так проста. Стандартная стратегия опирается на свойство степенной функции, с которым мы познакомились выше: гистограмма величины, соответствующей степенному распределению выглядит как прямая линия в двойных логарифмических координатах. Однако, простое построение гистограмм такого рода, и выяснение того, выглядят ли они при этом прямой линией в большинстве случаев – плохая практика.



*Рис 2. (a) Гистограмма набора 1 миллиона случайных чисел, описанного в тексте - они имеют степенное распределение с показателем α=2,5. (b) Та же самая гистограмма в двойных логарифмических координатах. Заметим, как зашумлена правая часть распределения. Это происходит из-за того, что число значений, попадающих в корзины в этой части распределения становится небольшим и статистические флуктуации становятся относительно большими. (с) Гистограмма, использующая "логарифмические корзины". (d) Кумулятивная гистограмма тех же данных, другое название "диаграмма ранг/частота". Кумулятивное распределение также имеет форму степенной функции, но с показателем α-1=1,5.*

Как мы видели выше, чтобы выявить степенную форму распределения лучше построить гистограмму в логарифмических координатах, и когда мы это сделаем для нашего набора данных, мы увидим характерную форму степенного распределения в виде прямой линии, рис.2 (b). Однако, этот график в некотором смысле не очень хорош. Конкретно, правая часть распределения зашумлена из-за статистических флуктуаций. В этой области степенное распределение истощается, так что в корзины попадает мало случаев (или их не попадает вовсе). Это приводит к тому, что относительные флуктуации количества случаев в корзинах становятся значительными, и возникает шум на диаграмме. Один из способов решить эту проблему – просто отбросить данные хвоста распределения. Но в этих данных часто содержится ценная информация и, более того, как мы увидим в разделе II.А, многие распределения соответствуют степенному распределению только в своём хвосте, так что есть опасность вместе с водой выплеснуть и ребёнка.

Альтернативное решение – варьировать размер корзин. Для этого мы сначала должны нормализовать число случаев по ширине корзин, в которые они попадают. То есть, число случаев в корзине шириной Δx должно быть поделено на Δx, чтобы получить число случаев на каждый единичный интервал величины x. После этого нормализованное число случаев в среднем перестаёт зависеть от ширины корзин и мы получаем возможность изменять ширину корзин по своему усмотрению. Чаще всего корзины устраивают таким образом, что каждая следующая шире предыдущей в определенное число раз - это так называемые логарифмические корзины. Для нашего случая, к примеру, мы можем выбрать множитель 2 и устроить корзины, которые перекрывают интервалы от 1 до 1,1, от 1,1 до 1,3, от 1,3 до 1,7 и т.д. (то есть, ширина корзин увеличивается: 0,1, 0,2, 0,4 и т.д.). Благодаря этому корзины в хвосте распределения получат больше вхождений, чем при фиксированной ширине корзин, и это уменьшит статистические ошибки в хвосте. Кроме того, мы получаем дополнительный полезный результат: на логарифмической шкале логарифмические корзины выглядят так, будто они имеют одинаковую ширину.

Именно логарифмические корзины мы применили для построения гистограммы на рис. 2(b), поэтому расстояния между точками, представляющими отдельные корзины, одинаковы на логарифмической шкале:

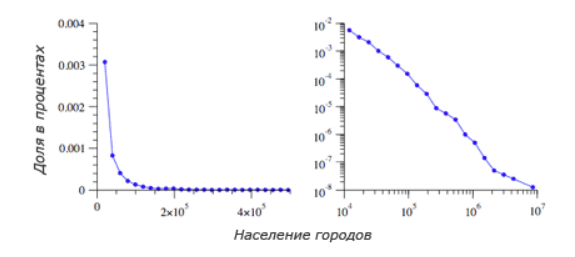
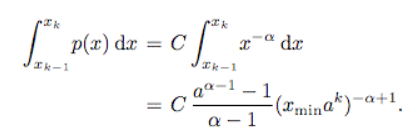


Рис.3 Логарифмирующая шкала

На рис.2(с) мы сделали то же самое для наших искусственно сгенерированных данных. Как видим, прямая линия степенного закона на этой гистограмме выглядит гораздо яснее, она просматривается по меньшей мере на порядок дальше, чем на рис.2(b).

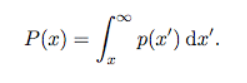
Но даже с логарифмическими корзинами в хвосте распределения остается некоторый шум (хотя его становится гораздо меньше). Предположим, что начало самой первой корзины находится в точке xmin, а отношение размеров соседних корзин равно a.

Тогда логарифмическая корзина с номером *k* захватывает промежуток от *xk−1 = xminak−1* до *xk = xminak*, и ожидаемое число случаев, которые попадут в эту корзину равно:

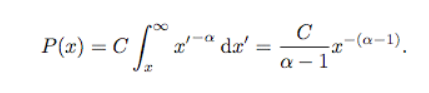


Таким образом, если *α > 1*, число случаев, попадающих в корзины уменьшается с ростом номера корзины *k*, и с приближением к хвосту распределения статистический шум усиливается. Как мы увидим в следующем разделе, большинство натуральных степенных распределений имеют показатель *α*, лежащий в промежутке между 2 и 3, поэтому шум в хвостах распределений - нормальное явление.

Другой, и во многих отношениях гораздо более предпочтительный способ анализа данных - вычисление *кумулятивной функции распределения*. Вместо того, чтобы строить простую гистограмму, мы строим диаграмму вероятности *P(x)* того, что что значение величины больше или равно некоторому *x*:



Диаграмма, которую мы получаем таким образом, не является простым представлением распределения данных, но она, тем не менее, весьма полезна. Если распределение соответствует степенному закону *p(x) = Cx−α*, тогда кумулятивная функция распределения *P(x)*:



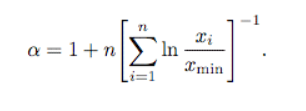
То есть, эта функция также соответствует степенному закону, но с показателем α − 1, что на единицу меньше исходного показателя. Таким образом, если мы построим диаграмму P(x) в двойных логарифмических координатах, мы тоже получим прямую линию, но с меньшим наклоном.

Заметим, что для построения P(x) нам не нужно раскладывать данные по корзинам. По своему определению, кумулятивная функция распределения задана для каждого конкретного значения x, и её можно строить как совершенно нормальную функцию без раскладывания по корзинам. Это снимает вопрос о том, какого размера должны быть корзины. Кроме того, это гораздо более бережный подход: раскладывание данных по корзинам усредняет, смешивает конкретные значения, попадающие в одну корзину, и мы теряем информацию, заключенную в индивидуальных значениях величины. Кумулятивное распределение не теряет информации индивидуальных значений - они все будут представлены на диаграмме.

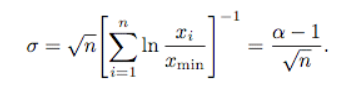
На рис.2(d) наши сгенерированные данные представлены в виде кумулятивного распределения, и мы вновь видим красноречивую прямую линию степенного закона, хотя и имеющую меньший наклон. Такого рода кумулятивные распределения ещё называют диаграммами ранг/частота по причинам, описанным в Приложении А. Говорят также, что кумулятивные распределения, соответствующие степенному закону, отвечают закону Зипфа или распределению Парето - в честь двух учёных, впервые начавших использовать этот тип распределений. Поскольку степенные кумулятивные распределения подразумевают и степенную форму обычного распределения p(x), "закон Зипфа" и "распределение Парето" - это по существу синонимы "степенного распределения". (Закон Зипфа и распределение Парето различаются в том, каким образом строится диаграмма кумулятивного распределения - Зипф использовал диаграммы, в которых x - на горизонтальной оси, а P(x) - на вертикальной. Парето поступал наоборот. Это приводит к множеству путаницы в литературе, хотя результаты, конечно, при этом не отличаются.

Мы знаем точное значение показателя α для нашего сгенерированного набора данных, поскольку сами его установили, но на практике приходится определять этот показатель для опытных данных. Один из способов это сделать - подобрать к представленному в двойных логарифмических координатах распределению наиболее близкую прямую линию, используя обычный метод наименьших квадратов. К сожалению, известно, что этот метод в случае аппроксимации степенной функции дает систематическую погрешность в определении показателя α, поэтому на него не следует полагаться. Например, аппроксимация методом наименьших квадратов прямой линии на рис.3(b) дает α = 2,26 ± 0.02, что явно расходится с известным нам точным значением α = 2,5.

Альтернативный простой и надёжный метод определения степенного показателя - использование формулы:



Тут xi , i= 1 ... n – опытные замеры величины x, а xmin - минимальное значение x. (Как будет показано в следующем разделе, на практике xmin обычно соответствует не самому малому опытному значению x, а самому малому значению, при котором ещё выполняется степенной закон.) Ожидаемая погрешность σ при использовании формулы равна:



Используя формулы (5) и (6) к нашему набору данных, мы получим оценку показателя *α = 2,500 ± 0,002*, что хорошо согласуется с известным точным значением 2,5.

## 2. Типы графиков степенных распределений.

В литературе по степенным законам часто используется три различных типа распределений:

1. Частотное степенное распределение

2. Ранговое степенное распределение

3. Кумулятивное степенное распределение

Одни и те же данные - например, данные по населению городов - можно представить в форме любого из трех распределений. Обычно, если мы используем один из трех типов и получаем в результате степенную кривую, то, используя любой другой тип мы тоже получим степенную кривую. Однако, от выбора типа распределения зависят конкретные параметры этой кривой и, в частности, показатель степени k. Например, частотное распределение городов по населению обычно соответствует степенному закону с показателем -2, а кумулятивное и ранговое распределение – степенному закону с показателем -1. Поэтому, встречая, например, где-то сообщение, что "распределение военных конфликтов по числу жертв соответствует степенному закону с показателем -1.8", нам всегда необходимо уточнять: какое именно распределение было использовано. Далее, чтобы не путаться, мы будем обозначать соответствующие каждому распределению показатели степени так:

1. Показатель частотного распределения: K(freq)

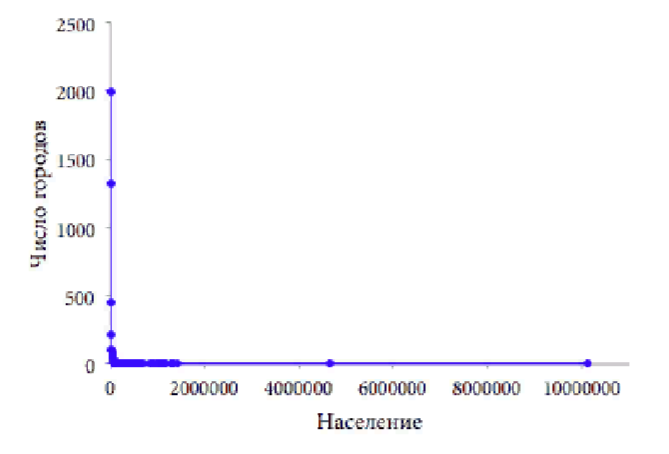
2. Показатель рангового распределения: K(rank)

3. Показатель кумулятивного распределения: K(cumm)

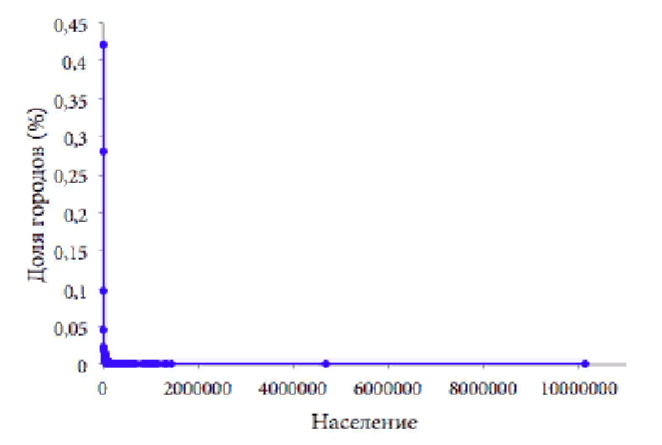
Давайте разберёмся с ними по очереди, опираясь на конкретный реальный пример.

### 2.1. Частотное распределение.

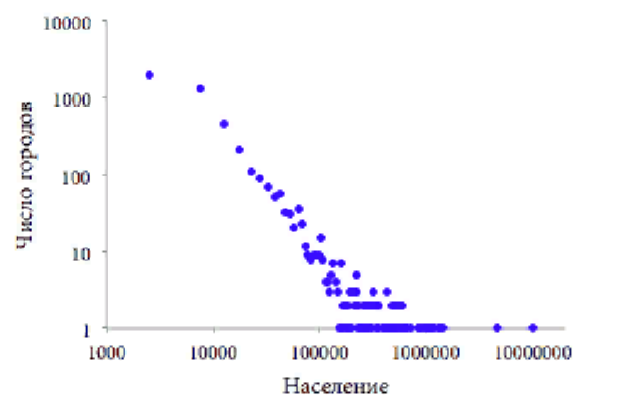
Частотное распределение - пожалуй, самое распространённое, но не потому, что оно удобнее или полезнее остальных, а по привычке: именно такой тип распределений традиционно применяется в статистической физике и теории вероятностей. Оттуда он перекочевал в статистику и теперь широко распространен под именем "гистограмма". Пусть, например, мы изучаем населённые пункты России с точки зрения их населённости. Мы взяли данные Госстата по 4718 населённым пунктам. Построим гистограмму. Для этого мы берём шкалу населённости и делим её на какие-то равные промежутки, "корзины". Мы можем, например, разделить шкалу на корзины по 5000 человек. В первую корзину мы складываем все населённые пункты, в которых живет от 0 до 5000 человек, во вторую - от 5000 до 10000 человек и т.д. Разложив по корзинам все города России, мы можем взглянуть на результат:



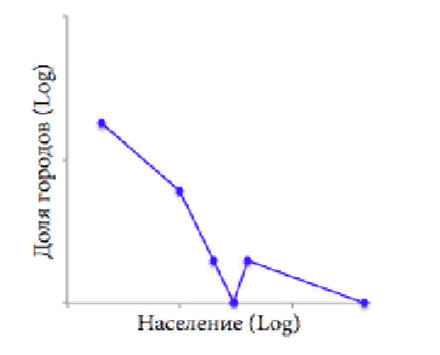
Обратим внимание, что по оси Y мы отмечаем просто количество городов, попавших в соответствующую корзину. Но мы могли бы отмечать не количество, а относительную долю, которую составляет содержание каждой корзины к общему числу городов - для этого поделим каждое значение шкалы Y на 4718 (столько у нас всего городов):



На такой модифицированной гистограмме видно, что в первую корзину (население от 0 до 5000 человек) попало 0,42 всех населённых пунктов, то есть, 42%. Мы можем также сказать, что какой-то конкретный населённый пункт России с вероятностью 42% окажется в первой корзине, поэтому такие распределения ещё называют распределениями плотности вероятности. Попробуем понять: подчиняются ли полученные данные степенному закону? Для этого, как мы уже поступали раньше, отобразим гистограмму в двойной логарифмической шкале:



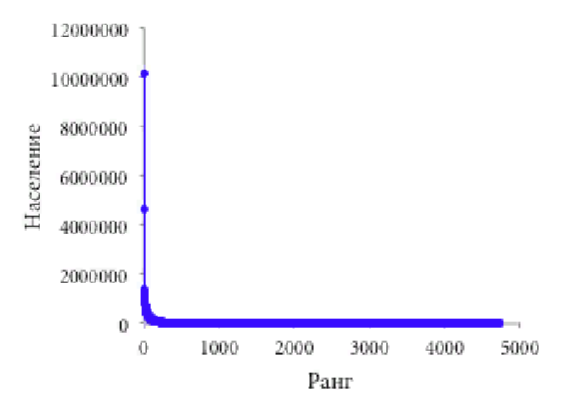
В целом прямая линия, вокруг которой группируются точки, вполне просматривается (существенно выпадают лишь крайние справа две точки, они соответствуют СанктПетербургу и Москве). Её наклон соответствует примерно K(freq)=-1,7 – -1,9 (точнее определить трудно). Однако, мы видим, что правая часть распределения зашумлена. Эта "борода" возникает из-за того, что в области высоких значений статистического параметра перестаёт действовать усреднение, которое эффективно сглаживает кривую в области низких значений. Попросту, в корзины, расположенные в начале шкалы попадает очень много городов и случайные вариации усредняются. А вот в корзины, расположенные в конце шкалы городов попадает мало и случайные вариации становятся очень заметными. Эта проблема становится критической, когда статистических данных не очень много. "Корзинное" построение распределения резко сокращает число значимых точек на диаграмме. Например, у нас 4718 городов легли всего в 105 корзин, а значит у нас осталось всего 105 значимых точек (не пустых корзин) на распределении. А если взять не все города России, а например, только города Калининградской области (их всего 30), и разложить по корзинам размером в 10000 человек, то вообще получается всего 6 значимых точек, по которым затруднительно распознать степенную функцию:



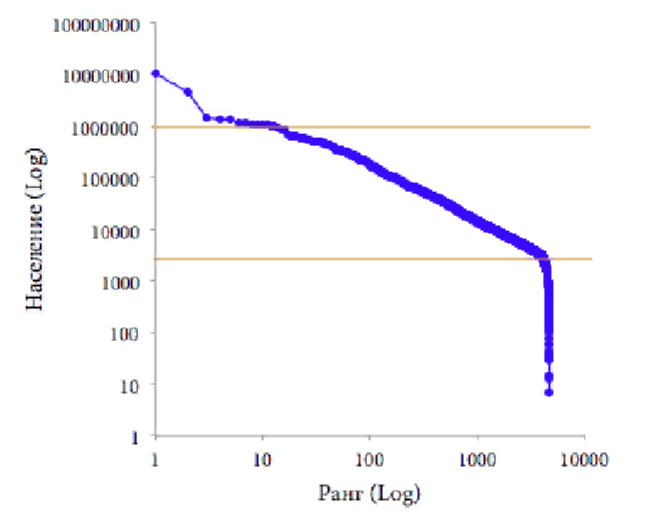
Эта проблема, а также относительная трудоёмкость построения частотного распределения (необходимо раскладывать объекты по корзинам) заставляет приглядеться к другому варианту представления степенной статистики – гораздо более "бережному" к исходным данным и простому в построении.

### 2.2. Ранговое распределение.

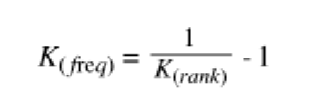
Этот тип распределений связывают с именем лингвиста Джорджа Зипфа, который в середине 20-го века обнаружил, что частота употребления слов в естественных языках соответствует степенному закону. Однако, в своём анализе Зипф использовал не частотное распределение, а ранговое. На нашем примере познакомимся с тем, как оно строится. Мы берём данные по населению городов и просто сортируем города в порядке убывания их населения. Номер, который получает в этом списке каждый город именуется его рангом. Теперь нам достаточно построить диаграмму, в которой по оси Y откладывается население каждого города, а по оси X - его ранг:



Соответствует ли это распределение степенному закону? Чтобы узнать, построим его в двойных логарифмических координатах:



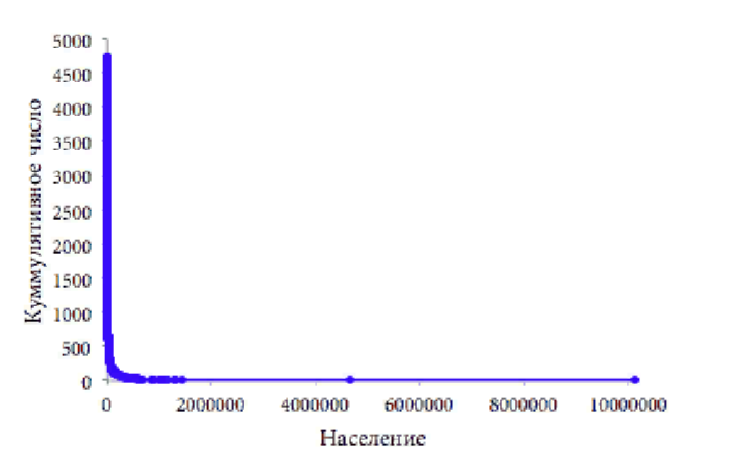
Мы видим, что для российских городов с населением более 3000 человек степенной закон хорошо выполняется, особенно в промежутке между 3000 и 1000000 человек (промежуток отмечен оранжевыми границами). Города-миллионники (не считая Москвы и Петербурга) выбиваются из общей картины, они существенно "не добирают" населения. Резкое искажение общей картины имеется и для населённых пунктов с населением меньше 3000 человек - их будто бы "слишком мало". Впрочем о том, какой смысл имеет прямолинейность общей линии и что значат отклонения от неё мы ещё будем говорить. Огромное преимущество рангового степенного распределения в том, что значимыми у нас остаются все имеющиеся данные, они все представлены точками на распределении. Благодаря этому существование степенного закона, а также показатель степени можно установить с гораздо большей точностью, нежели при частотном распределении. В данном конкретном примере показатель степени K(rank) оказывается равным -1,09 (с довольно высокой точностью). Мы говорили, что если в статистике явления имеется степенной закон, то он будет проявляться на любом из трёх типов распределений, с которыми мы знакомимся. Но показатели степени будут во всех трёх случаях разные. И вот, на нашем примере мы видим, что при использовании частотного распределения мы получили показатель степени около -1,8, а при ранговом - показатель -1. Сравнивая эти два значения, можно подумать, что между ними разница составляет около 1. Так и есть, можно доказать теоретически (мы тут опустим доказательство, его можно найти, например, тут), что за исключением редких случаев между показателем частотного степенного распределения K(freq) и показателем рангового K(rank) действует следующее соотношение:



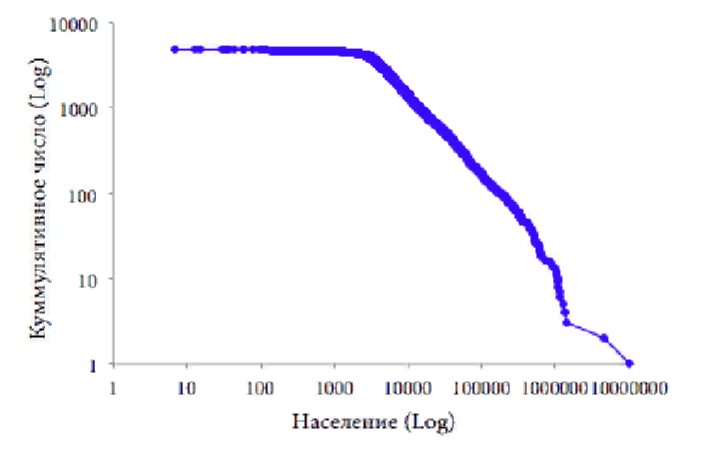
В нашем случае K(rank)=-1,09, значит, K(freq) должен быть равен примерно -1,92. Это соответствует нашему приблизительному практическому результату (-1,7 – -1,9) и разница между K(freq) и K(rank), действительно, составляет около единицы.

### 2.3. Кумулятивное степенное распределение.

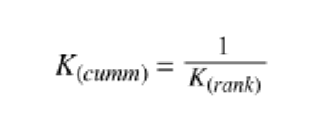
Близким родственником рангового распределения является кумулятивное распределение или, как еще его называют, распределение Парето. Так говорят в честь итальянского экономиста Вильфредо Парето, который в начале 20-го века заметил, что 80% землевладений в Италии находятся в руках всего лишь 20% населения. На кумулятивном распределении по оси X отмечается величина параметра, у нас это население города, а по оси Y - количество городов, население которых больше или равно текущему X. Скажем, в нашем примере для точки X=909341 (население Красноярска) получается Y=14, потому что в России есть только 14 городов, население которых превышает или равно 909341 человек. Следующей точкой будет X=1001653 (Пермь), и значение Y для этой точки равно 13. Последней точкой распределения будет Москва (X=10126424, Y=1). Построим кумулятивное распределение для нашего примера:



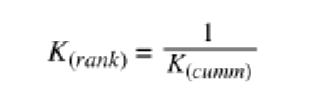
И оно же в двойных логарифмических координатах:



Наблюдательный читатель может заметить, что его форма очень похожа на форму рангового распределения - только перевёрнутого, словно оси X и Y поменялись местами. И это не случайно. Действительно, кумулятивное распределение является ничем иным, как обращенным ранговым распределением. И просто понять, почему: последняя точка на кумулятивном распределении соответствует Москве (X=10126424, Y=1). Но Москва же оказывается и первой точкой рангового распределения (X=1, Y=10126424). Далее, Санкт-Петербург - это предпоследняя точка кумулятивного распределения (X=4661219, Y=2), но вторая точка рангового распределения (X=2, Y=4661219). Получается, что движение от конца кумулятивного распределения к его началу в точности соответствует движению по ранговому распределению, но наоборот, от начала к концу. Трудно сказать, почему Парето в своих работах предпочёл кумулятивное распределение более простому и понятному ранговому. Вероятно, тут сыграло свою роль, что в нём по оси X откладываются значения статистического параметра (у нас это население) - также как в "научно правильном" частотном распределении. Как бы то ни было, кумулятивные распределения получили широкое распространение и нам важно их не путать ни с частотными ни с ранговыми. кумулятивное распределение является обратным ранговому и это позволяет нам легко установить соотношение между их показателями степени: показатель степенного кумулятивного распределения точно обратен показателю степенного рангового распределения:



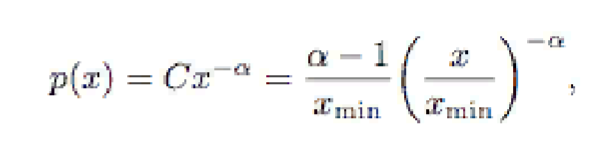
Разумеется, и наоборот:



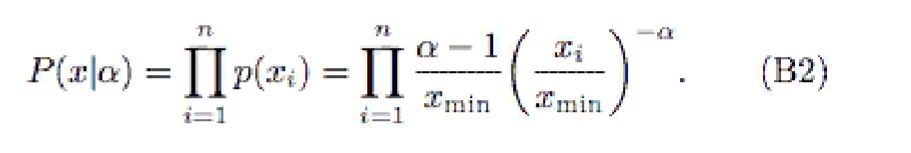
Конкретно, в нашем примере получается, что показатель кумулятивного распределения (распределения Парето) K(cumm) равен 1/-1,09 = -0,92 Итак, мы теперь можем записать парные соотношения между показателями степенных распределений трёх типов.

## 3. Определение параметра степенного распределения методом максимального правдоподобия.

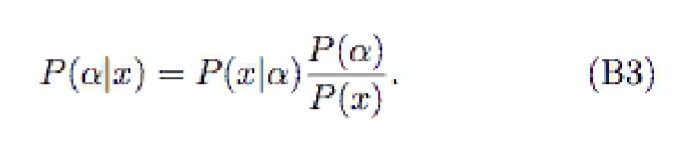
Рассмотрим степенное распределение:



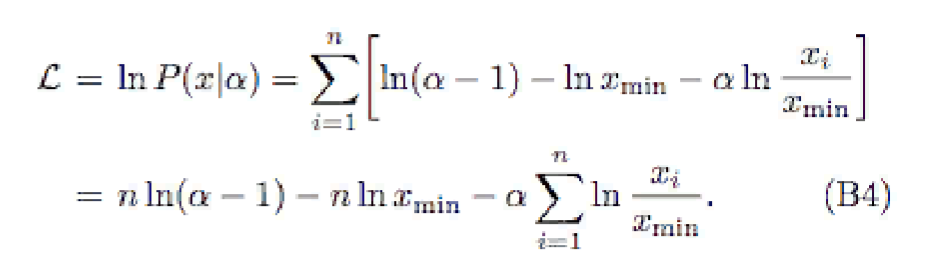
в котором используется значение нормирующей константы C, вычисленное в уравнении (9). Если у нас имеется набор из n значений xi, вероятность, что эти значения получены из данного распределения пропорциональна величине



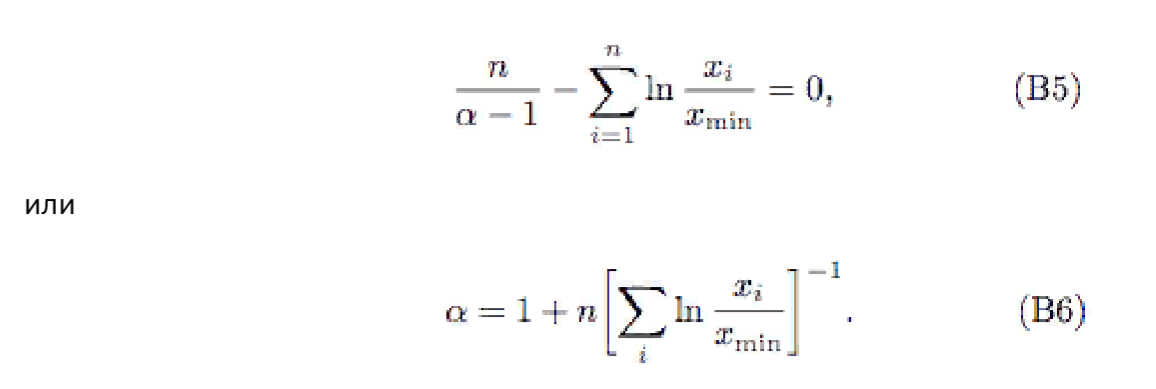
Она именуется правдоподобием набора данных. Чтобы найти значение α, которое лучше всего соответствует данным, необходимо вычислить вероятность P(α|x) некоторого значения α исходя из набора данных {xi}. Эта вероятность связана с P(x|α) законом Байеса, так что:



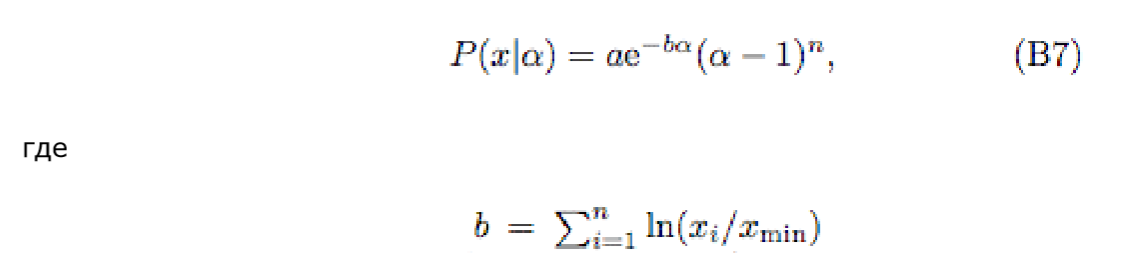
Исходная вероятность данных P(x) фиксирована, поскольку x само фиксировано - x это конкретный набор наблюдений у нас на руках и он не меняется при расчетах. Кроме того, если нет каких-либо свидетельств об обратном, обычно предполагается что исходная вероятность степенного показателя P(α) однородная, то есть, является константой, независимой от α. Таким образом P (α|x) пропорциональна P (x|α). Удобнее работать с логарифмом P(α|x), который с точностью до аддитивной константы равен логарифму правдоподобия. Обозначим его как L:\



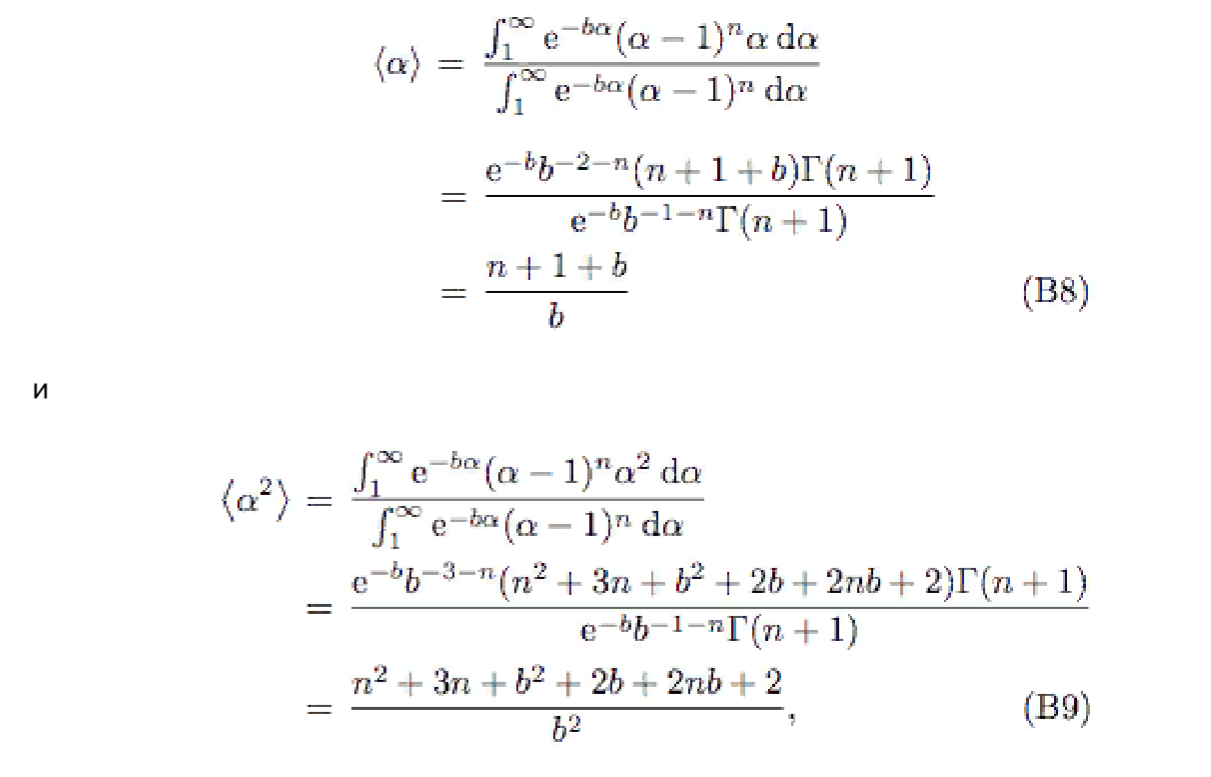
Теперь посчитаем наиболее вероятное значение α, максимизируя правдоподобие относительно α. Это то же самое, что максимизирование логарифма правдоподобия, поскольку логарифм - монотонно возрастающая функция. Устанавливая ∂L/∂α = 0, мы найдем, что



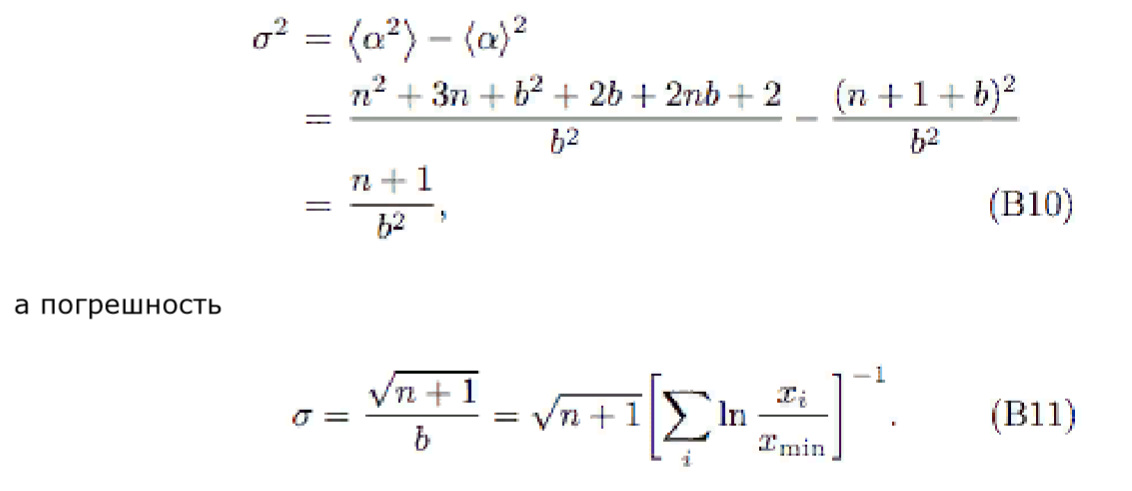
Нам также нужно узнать ожидаемую погрешность значения α. Мы можем оценить её из ширины максимума правдоподобия как функции от α. Взяв экспоненту от уравнения (B4), мы выясним, что правдоподобие имеет форму



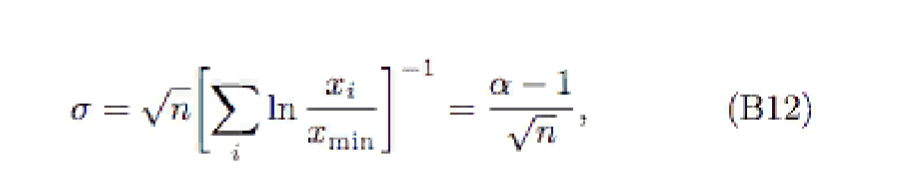
тут a - константа нормировки, не представляющая интереса. Предположив, что α > 1, так что распределение (B1) поддается нормировке, среднее и средний квадрат величины α в этом распределении задается выражениями



тут Γ(x) - гамма-функция. Тогда стандартное отклонение α равно



В большинстве случаев n велико, поэтому можно считать, что n+1 ≈ n, при этом получим



где α - оценка показателя, полученная методом максимального правдоподобия по уравнению (B6).

## 4. Применение теста Колмогорова-Смирнова для сравнения теоретической и экспериментально полученной функции распределения.

Данный критерий позволяет оценить существенность различий между двумя выборками. Его применение возможно также для сравнения эмпирического распределения с теоретическим.

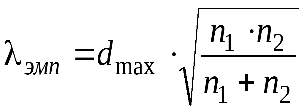
Объёмы рассматриваемых выборок должны быть достаточно большими: ≥50, ≥50. Для использования теста выборки должны быть представлены в виде частотного распределения, при этом число категорий должно быть небольшим (до 7-9).

Нулевая гипотеза H0={различия между двумя распределениями недостоверны}.

Критерий позволяет найти категорию, в которой сумма частот расхождений между двумя распределениями является наибольшей, и оценить достоверность этого расхождения.

Алгоритм проверки:

1. Определяются категории значений признака.
2. Строится частотное распределение каждой выборки по выделенным категориям.
3. Вычисляются относительные частотыhttps://studfile.net/html/2706/304/html__AlxJSSZga.0n5C/img-45iWX6.png, равные частному от деления частот на объём выборки, для каждой из имеющихся выборок.
4. Определяется модуль разности соответствующих относительных частот.
5. Определяется наибольший модуль, который обозначаетсяhttps://studfile.net/html/2706/304/html__AlxJSSZga.0n5C/img-PCpkAX.png.
6. Вычисляется эмпирическое значение критерия https://studfile.net/html/2706/304/html__AlxJSSZga.0n5C/img-DizEv3.png:



1. Определяется критическое значение критерия для выбранного уровня значимости.
2. Если эмпирическое значение критерия больше критического, то нулевая гипотеза отвергается, и группы по рассмотренному признаку отличаются существенно.

Схематично алгоритм применения критерия Колмогорова-Смирнова можно представить следующим образом:

