UNIVERSIDADE ESTADUAL PAULISTA
Instituto de Geociências e Ciências Exatas
Campus de Rio Claro

DISTRIBUIÇÕES ESTATÍSTICAS: CITAÇÕES DE PUBLICAÇÕES CIENTÍFICAS E DE CIENTISTAS

Rosana Angélica Gonçalves Pesce

Orientador: Prof. Dr. Hari Mohan Gupta

Dissertação de Mestrado elaborada junto ao Programa de Pós-Graduação em Física – Área de Concentração em Física Aplicada, para obtenção do título de Mestre em Física.

Rio Claro (SP)

2005

UNIVERSIDADE ESTADUAL PAULISTA
Instituto de Geociências e Ciências Exatas
Campus de Rio Claro

DISTRIBUIÇÕES ESTATÍSTICAS: CITAÇÕES DE PUBLICAÇÕES CIENTÍFICAS E DE CIENTISTAS

Aluna: Rosana Angélica Gonçalves Pesce

Orientador: Prof. Dr. Hari Mohan Gupta

Comissão Examinadora

Prof. Dr. Hari Mohan Gupta – Orientador
IGCE/UNESP/Rio Claro

Prof. Dr. Gerson Antonio Santarine
IGCE/UNESP/Rio Claro

Profa. Dra. Rosane Riera Freire
PUC/RJ

Rio Claro, 26 de outubro de 2005.

Resultado:Aprovada

DEDICATÓRIA

À Deus, aos meus pais, à minha família,
ao meu esposo e aos que virão.

AGRADECIMENTOS

Ao Prof. Dr. Hari Mohan Gupta, meu orientador, pela paciência, carinho e atenção dispensada para a realização não só de um trabalho, mas de um sonho...

Ao Prof. Dr. Gerson Antonio Santarine e à Profa. Dra. Rosane Riera Freire, pelas sugestões que me ajudaram a aprimorar este trabalho.

À todos os docentes e funcionários do Departamento de Física, em especial ao chefe, Prof. Dr. Alberto Ibañez Ruiz, que possibilitou com que eu me dedicasse aos estudos.

Ao Prof. Dr. José Roberto Campanha pelos “toques” na utilização do Mathematica.

Aos Profs. Drs. Jorge Roberto Pimentel e Dimas Roberto Vollet, coordenadores do Programa de Pós-Graduação em Física no período em que fui aluna, por permitirem com que eu participasse deste programa.

À Maristela, pela colaboração e apoio nos momentos em que me ausentei para o cumprimento das disciplinas.

Aos meus amigos da AFA, que nunca me deixaram desistir.

À toda minha família, pelo incentivo.

Aos meus pais, exemplos de dedicação, cujos maiores ensinamentos foram a prática da humildade e o gosto pela aquisição do conhecimento.

Ao meu amiguinho Billy, pela paciência...

Ao meu esposo, pelo amor e por acreditar em mim.

Obrigada.

SUMÁRIO

ÍNDICE

ÍNDICE DE TABELAS

ÍNDICE DE FIGURAS

RESUMO

O número de vezes que um cientista ou uma publicação científica é citada em outra publicação científica é um fator importante na consideração do seu mérito. No presente trabalho, estudamos a distribuição estatística do índice de citação de publicações científicas e de cientistas. Estudamos a densidade da publicação científica versus a citação. Encontramos que a estatística de Tsallis (Lei de Potência) pode explicar toda a distribuição acima de 8 graus de magnitude (10^{-4} a 10^{4}). Confirmamos isto através do gráfico de Zipf. Normalmente os índices de citação dos cientistas mais citados são acessíveis, o que em troca dão somente a informação sobre a dinâmica do mecanismo de citação deste grupo. Estudamos a distribuição estatística do índice de citação de físicos e químicos, brasileiros e internacionais, através do Gráfico de Zipf. Como os cientistas brasileiros são um pequeno sub-grupo no contexto da comunidade científica internacional, isto pode melhor explicar a dinâmica do índice de citação. Sendo assim, encontramos que a distribuição de lei de potência gradualmente truncada é a que melhor explica a distribuição do índice de citação com quase os mesmos valores dos parâmetros. Finalmente, discutimos o possível mecanismo por trás do índice de citação de cientistas e publicações científicas.

Palavras chave: Lei de Potência; citação; publicação.

ABSTRACT

The number of times, a scientist or a scientific publication is cited in other scientific publication is now an important factor in his merit consideration. In the present work, we studied the statistical distribution of the citation index of scientific publications and scientists. We studied the scientific publication density versus citation. We find that Tsallis (Power Law) statistics can explain the entire distribution over eight orders of magnitude (10^{-4} to 10^{4}). We further confirm it through Zipf plot. Normally citation indices of highly cited scientists are available, which in turn give information only about the dynamics of the citation mechanism of this group. We studied the statistical distribution of the citation index of Brazilian and international physicists and chemists, through Zipf-plot technique. As Brazilian scientists are a small sub-group within the international scientific community, it can better explain the dynamics of citation index. We find that the gradually truncated power law distribution explain well distribution of citation index with almost same parameter values. We finally discuss possible mechanisms behind citation index of scientists and scientific publications.

Key words: Power Law; citation; publication.

Capítulo 1

Introdução

1.1. O Método Científico

O método científico começou a ser elaborado por pensadores da Antigüidade, sendo mais tarde aperfeiçoado e adotado, por unanimidade, pelos pesquisadores modernos.

Apoiado na observação, isto é, procurando explicar os processos da natureza através de experimentos que uma vez demonstrados podem ser repetidos por

qualquer outro pesquisador, o método científico consta de pelo menos quatro etapas consecutivas (Einstein, 1966; Macau, 2002; Meadows, 2000):

1) Identificação do problema com pesquisa bibliográfica

Quando se estuda algum tema, é porque este surgiu de algum problema que se quer resolver ou entender melhor. A identificação deste problema, que deve ser enunciado claramente e descrito de modo que todos possam compreender, é a primeira etapa do método científico. Além disso, o pesquisador também faz uma pesquisa bibliográfica na área do tema escolhido, para se ter uma visão prévia dos trabalhos já publicados sobre o assunto.

2) Experimentos práticos e simulação computacional

A segunda etapa é geralmente a mais longa, uma vez que consiste na realização de experimentos práticos e modelagens teóricas, que visam elucidar ou resolver tal problema. Quando os estudos estiverem avançados a ponto de haver uma nova ou mais detalhada explicação para o problema, é que iremos para a terceira etapa.

3) A publicação dos resultados e a divulgação do trabalho

Caso os estudos e experimentos já tenham chegado a algum novo entendimento, estes devem ser descritos com clareza e, submetidos à comunidade científica, que irá analisar o manuscrito podendo recomendar (talvez com sugestão para modificações), ou rejeitá-lo para publicação. No caso de ser recomendado, o trabalho é publicado numa revista adequada para o assunto, podendo assim, ser lido e citado por outros pesquisadores. A publicação dos resultados de uma pesquisa é a principal finalidade da carreira de um cientista, indicando a primazia de idéias, permitindo sua livre discussão, além de contribuir para o avanço da atividade científica.

A publicação de um trabalho pode ser feita em uma revista especializada, Anais de Conferências ou Congressos Científicos, ou mais recentemente na Internet no formato “on-line”. Em Ciências Sociais, algumas vezes, os trabalhos são publicados na forma de livros, o que não acontece em Ciências Exatas ou Biológicas.

Tanto em periódicos especializados como em congressos, a publicação científica é a forma mais tradicional de se apresentar os resultados de pesquisas e colocá-los à disposição da sociedade interessada (universitários, pesquisadores, estudantes, etc). É, portanto, de fundamental importância que haja publicações, caso contrário, os trabalhos que estão sendo efetuados não teriam mecanismos para difusão do conhecimento adquirido. É fato que, a publicação é a maneira internacionalmente consagrada de avaliação técnica da qualidade da pesquisa.

A produção científica adquire maior importância quando se considera que este é um indicador utilizado pelas agências de fomento para concessão de verbas destinadas à pesquisa. Porém, deve-se ressaltar que alta produtividade nem sempre é sinônimo de alta qualidade.

4) A citação do trabalho

Uma vez que o trabalho foi publicado, este é lido e analisado por inúmeros pesquisadores que trabalham na mesma área ou afins. Se o trabalho não apresenta importância ou é de rotina, geralmente é ignorado. Quando contém alguma novidade, em geral é criticado favoravelmente ou não dependendo do assunto, e, dessa maneira, tem como conseqüência a sua citação. Se a crítica é desfavorável o trabalho é esquecido depois de algum tempo. Porém, se a crítica é positiva e nos proporciona nova teoria ou linha de pesquisa, o trabalho começa a ser cada vez mais citado, confirmando ou modificando esta nova teoria. Caso a teoria seja confirmada, com o tempo torna-se uma lei. Se houver algum resultado contrário, a teoria necessita ser modificada ou até mesmo rejeitada, abrindo-se espaços para novas teorias.

No processo de pesquisa, existe sempre um assunto que é revisado por algum outro cientista da área e publicado como trabalho de revisão em revistas especializadas, tais como: a Advances in Physics, Review of Modern Physics, entre outras da área de Física. Algumas vezes estes trabalhos de revisão também são publicados em livros.

1.2. O Índice de Citações

Milhares de artigos são publicados em periódicos científicos todos os anos, e todos contêm listas de citações. O índice de citações de um trabalho é determinado pelo número de vezes que uma publicação é referenciada por outros autores. Este indicador serve para demonstrar o impacto de determinado estudo perante a comunidade científica, além de avaliar a quantidade de trabalho do pesquisador.

Em geral, uma publicação científica é primeiramente citada por pessoas que trabalham no mesmo grupo porque estão familiarizadas com o trabalho. Mais tarde, isto atrai a atenção de outros grupos que trabalham na mesma área e, logo alguns artigos importantes se tornam clássicos em determinada área de pesquisa e são citados somente para completar a introdução do problema, embora muitas vezes o autor realmente não leia o mesmo e o seu problema seja um tanto diferente. Ao mesmo tempo em que a maioria dos artigos é esquecida nos primeiros anos, estes importantes artigos são citados por um longo tempo.

Embora, alguns gostem e outros não, o índice de citação indica o prestígio de seu trabalho, este porém, deve ser usado com cautela.

1.3. O Acervo Científico do ISI

Circulam hoje pelo mundo milhares de periódicos. Aproximadamente 8.500 encontram-se indexados no banco de dados do Institute for Scientific Information (ISI), situado na Filadélfia, EUA. Apesar do ISI trabalhar com publicações em 36 línguas diferentes, a maioria dos jornais, particularmente nas áreas de Ciências Exatas e Biológicas, listados pelo instituto é redigida em inglês (Meadows, 2000).

Apesar de toda a crítica aos critérios de seleção dos títulos indexados, os indicadores produzidos pelo ISI são universalmente aceitos pela comunidade científica como indicadores de excelência na ciência. Os periódicos são indexados nessa base após rigorosa seleção, o que qualifica essa fonte de dados como uma das mais conceituadas do mundo (Santos, 2003). Também são considerados de

extrema valia para a ciência da informação, pois possibilitam monitorar o desenvolvimento da ciência sob a perspectiva das relações entre o avanço da ciência e da tecnologia e o progresso econômico e social (Zimba e Muller, 2004).

Neste trabalho, utilizamos os dados do ISI para as publicações e citações de físicos e químicos internacionais.

1.4. Proposta deste Trabalho

O objetivo deste trabalho consiste em analisar a distribuição estatística do índice de citação de físicos e químicos brasileiros e internacionais, além da densidade de publicação científica em comparação à citação. Os índices internacionais foram extraídos do ISI (Institute for Scientific Information) e os nacionais da Folha de São Paulo do ano de 1999. A organização deste estudo foi elaborada conforme se segue:

No Capítulo 2, discutir-se-á sobre sistemas complexos e a sua relação com a distribuição de publicação e citação.

O Capítulo 3, está destinado a uma breve descrição relativa às principais distribuições estatísticas de grande importância na discussão do tema proposto.

No Capítulo 4, abordaremos a distribuição de publicação científica em relação à densidade de publicação e a citação.

A distribuição de citação de físicos e químicos brasileiros e internacionais está descrita no Capítulo 5.

Para finalizar, no Capítulo 6, discutiremos o possível mecanismo por trás do índice de citação de cientistas e das publicações científicas.

Capítulo 2

Sistemas Complexos

2.1. As leis básicas da Física

“Começando pelo Big Bang, o Universo supostamente evoluiu de acordo com as leis da Física.” (Bak, 1997)

O século XIX e XX presenciou o apogeu do método científico, associado ao enfoque reducionista, em que os sistemas devem ser observados sob um nível crescente de resolução na busca por seus constituintes elementares. Em

decorrência disto, a matéria foi considerada como uma formação sucessiva de moléculas, átomos, núcleos, e quarks, constituindo-se a base de tudo o que existe na Natureza.

Sob o prisma do enfoque reducionista, encontramos as seguintes leis da natureza em Física:

a) Leis do Movimento de Newton;
b) Equações de Maxwell para sistemas elétricos e magnéticos;
c) Equações da Relatividade de Einstein para altas velocidades;
d) Mecânica Quântica para descrição de movimento e energia de partículas como elétrons, etc.

Estas leis poderiam ser escritas por equações em poucas páginas, mas, a Matemática envolvida para resolver tais equações da Física é extremamente complexa, principalmente envolvendo sistemas com três ou mais corpos. Por exemplo, calcular o movimento de dois planetas na presença de outros planetas e do Sol, requer cálculos extremamente laboriosos.

Mas o nosso mundo não é composto apenas por sistemas que podem ser definidos sob o enfoque reducionista para previsão de comportamentos futuros. Os terremotos, por exemplo, são um caso em que teoricamente poderíamos entender seu comportamento, desde que atribuíssemos condições iniciais a cada partícula elementar, medindo-se posições e velocidades associadas a trilhões e trilhões dessas partículas e depois computar individualmente a trajetória e o estado de cada uma delas inserindo-as num sistema de equações diferenciais de ordem extremamente elevada, num esforço impraticável na busca por soluções numéricas.

A termodinâmica por sua vez, introduziu novos conceitos de modelos e previsões utilizando-se apenas das variáveis que se mostrem relevantes para uma descrição satisfatória do comportamento de um sistema, ou seja, a utilização de variáveis macroscópicas para descrever satisfatoriamente o comportamento do sistema. Como no caso dos gases, através da utilização das variáveis macroscópicas de estado, pressão, volume e temperatura, é possível descrever ou predizer seu estado de equilíbrio, por meio da “lei dos gases perfeitos”. Esta descrição permite a utilização de um modelo simples e tratável, no nível

microscópico, possibilitando também fazer previsões do comportamento futuro, associadas ao comportamento do sistema como um todo, no nível macroscópico.

Somente em meados do século XIX, com a introdução da mecânica estatística por Maxwell, Boltzman e Gibbs, houve alteração no *conceito laplaciano de predição* (de que a partir do estado completo do universo num determinado momento, descrito pelas posições e velocidades de todas as partículas, é possível predizer todos os estados futuros), que se mostrava ineficaz em vários casos, assim como no exemplo da previsão do comportamento dos gases ou terremotos citados anteriormente. Neste caso, aplica-se o conceito de *predição probabilística*, calculando-se apenas a distribuição de probabilidade das variáveis e não a solução exata associada ao conjunto de condições iniciais.

Para descrever um sistema através da mecânica estatística, faz-se necessário aceitar que todas as possíveis combinações associadas aos processos rápidos, ou seja, processos que são muito mais velozes do que a escala de tempo de nossas observações, ocorrem de acordo com uma distribuição fixa de probabilidade. Assim, por meio de uma equação de movimento, calcula-se a distribuição de probabilidade associada a esta equação de movimento e suas propriedades. Conseguimos descrever na mecânica estatística, o comportamento de cristais que possuem regularidade periódica ou os gases, onde todas as partículas são estatisticamente iguais. Estes exemplos citados acima se enquadram na definição de **sistemas simples**, que poderia ser reduzida aos seguintes critérios:

- leis básicas usando equações matemáticas;
- análise de poucas partículas ou corpos;
- poucas interações;
- e possuem um estado de equilíbrio constante.

Felizmente, o mundo em que vivemos não é formado somente de gases e cristais. A Terra possui rios, montanhas, mar, rochas, etc; nossa história possui registros de guerras, revoluções, religiões, entre outros acontecimentos; a Economia é composta por consumidores, produtores, governos, etc; todos exemplos de sistemas com estruturas muito variadas. Além disso, todos estes sistemas se interagem constantemente, o que torna impossível a simples utilização das leis

básicas da Física para o estudo e entendimento das propriedades macroscópicas. Estes sistemas são denominados **sistemas complexos**, que não podem ser reduzidos em poucas partículas e interações. Além disso, esses sistemas não possuem estado de equilíbrio.

O termo **complexidade** vem do latim *complexus,* que significa entrelaçado ou torcido junto. A palavra **complexo,** segundo o dicionário Houaiss (2001): *diz-se, de um conjunto, tomado como um todo mais ou menos coerente, cujos componentes funcionam entre si em numerosas relações de interdependência ou de subordinação, de apreensão muitas vezes difícil pelo intelecto e que geralmente apresentam diversos aspectos.*

A computação deu aos cientistas a possibilidade de compreender melhor a dinâmica de sistemas tais como: o cérebro humano, a economia, uma colméia, um bando de andorinhas e até a eleição num país. Todos estes sistemas têm grupos de elementos distintos (neurônios, empresas, abelhas, pássaros, eleitores) que ao exercitarem motivações individuais, acabam produzindo efeitos característicos de algo maior (sistema), algo que não pode ser deduzido a partir do comportamento de cada elemento considerado isoladamente, mas que surge das interações entre eles.

O número de variáveis interdependentes nos sistemas complexos é extremamente grande e, com certeza, conhecemos apenas uma pequena parte delas.

Ainda que conheçamos tudo sobre os componentes de um sistema complexo, não implica que conhecemos o comportamento do sistema, podemos citar como exemplo o cérebro, que é composto por um número grande de neurônios que apesar de serem células complexas, possuem um comportamento simples, o da emissão ou não de um impulso elétrico. Apesar desse comportamento simples, a união de bilhões de neurônios interagindo entre si, faz com que apareçam propriedades complexas como os pensamentos.

Entender o comportamento de sistemas complexos significa compreender como suas diversas partes agem em conjunto de forma a produzirem o

comportamento do todo, em decorrência disso surgem os fenômenos coletivos e as propriedades que não estão presentes nas partes quando analisadas separadamente.

Estes sistemas são estudados através de métodos estatísticos e, portanto, não produzem detalhes específicos de cada partícula do sistema, apenas previsões sobre o comportamento do sistema como um todo.

O estudo de propriedades macroscópicas dos sistemas complexos fez surgir novas linhas de pesquisa em diversas áreas como: Física, Biologia, Geografia, Geologia, Economia, etc., que se utilizam de conceitos próprios com a finalidade de melhor estudar estes comportamentos.

O estudo de publicações de artigos científicos enquadra-se num caso particular de sistemas complexos, onde fatores como tamanho da universidade, infra-estrutura, equipamentos, laboratório, apoio técnico, bibliotecas, recursos financeiros, dentre outros, são determinantes na produção científica. Fatores que ao interagirem entre si, podem produzir ou não um alto índice de publicações.

No caso das citações, objeto deste trabalho, também estamos tratando com um sistema complexo, pois a interação entre os diversos cientistas do mesmo grupo que citam um mesmo artigo, chama a atenção de outros pesquisadores da mesma área de pesquisa e estes citam os mesmos artigos.

Uma tentativa para se definir sistemas complexos poderia ser a seguinte:

Sistemas complexos são sistemas nos quais as interações entre os componentes do sistema produzem propriedades denominadas emergentes ou coletivas, as quais não podem se deduzidas a partir das propriedades individuais dos seus componentes. Eles se auto gerenciam, isto é, não há controle central, o resultado final é conseqüência da interação dos elementos um com os outros. Ainda que independentes, os elementos produzem bolsões de cooperação, formando grupos ou comunidade que geram comportamentos sofisticados que nenhum agente individual produziria sozinho.

Sistemas complexos têm que aprender, ou seja, estes devem modificar-se de acordo com as condições oferecidas pelo meio, e é por isso que eles se auto gerenciam. Todo sistema complexo aprende através de realimentação com o meio exterior, incorporando em sua estrutura informações sobre seu meio exterior. Esses sistemas são adaptativos. Á medida que as condições externas mudam, a estrutura do sistema muda junto, automaticamente. Esta característica é particular dos sistemas adaptativos complexos.

O auto gerenciamento e o aprendizado através da realimentação tornam esses sistemas extremamente flexíveis. Grupos de elementos antes especializados em certas atividades desaparecem, e novos nichos são criados à medida que o ambiente muda. Assim, os agentes nunca ficam presos a comportamentos que foram úteis no passado, mas que ficaram obsoletos. É isso que faz o sistema como um todo se adaptar a mudanças.

2.2. Características dos sistemas complexos

Já nos reportamos anteriormente que ao se trabalhar com sistemas complexos, estuda-se basicamente a probabilidade estatística da ocorrência de um determinado comportamento e, portanto, não se consegue mostrar detalhes específicos de cada componente.

Alguns exemplos podem ser considerados conforme se segue:

- Podemos estudar a probabilidade de roubos de carro para uma seguradora, mas não podemos saber se determinado carro será roubado.

- Não é possível determinar a velocidade e posição de uma molécula de um determinado gás, porém, podemos definir uma distribuição de velocidade para tais moléculas.

- No caso do vestibular, é possível estudar a distribuição estatística das notas dos alunos de determinado colégio, mas é impossível saber se determinado aluno terá nota maior ou menor que 5.

- Nas citações de artigos científicos, podemos verificar a distribuição de citação dos físicos, mas não podemos prever se determinado artigo será mais ou menos citado.

Através de observações empíricas em várias áreas do conhecimento, detectamos uma propriedade de variância em escala em alguns sistemas que geralmente chamamos de sistemas complexos.

Esta propriedade possui as seguintes características:

2.2.1. Lei de Potência

Devido à interação entre seus elementos, os sistemas complexos podem exibir um comportamento catastrófico, onde uma parte do sistema pode afetar outra, em um efeito dominó. Neste contexto, pela Lei de Potência, a probabilidade de ocorrência de um evento x é proporcional a alguma potência de x, isto é, $P(x) \cong x^{-n}$.

Esta distribuição é devida à **realimentação positiva** (**feedback positivo – "bola de neve"**), em que quando acontece um evento, há um outro elemento que ajuda a aumentar a magnitude deste evento. Este tipo de estrutura cíclica também pode ser chamado de **laço de realimentação** (Capra, 1996).

Exemplos de **laços de realimentação**:

- quando um aluno apresenta alguma habilidade em determinada área, os pais e professores dão mais condições para que esse aluno aumente os seus conhecimentos nesta área, e assim, o mesmo acaba se sobressaindo dos demais colegas, o que chama a atenção do meio externo que lhe dá maiores incentivos como bolsas, estágios, etc.

- os clientes satisfeitos com um determinado produto contam para os amigos, que os compram e ficam satisfeitos, e então fazem propaganda para seus amigos...

- em uma briga ou discussão, alguém faz algo que provoca uma resposta do outro, e esta provoca nova resposta do primeiro, e assim por diante.

- nos motores de combustão interna (motores de automóveis e motos), a rotação do motor é que provoca a alimentação de combustível, e por isso ele precisa iniciar sua rotação para que o ciclo alimentação-explosão-rotação-alimentação funcione.

Com relação às publicações científicas, sabemos que um artigo que inicialmente é citado mais vezes atrai a atenção de mais pesquisadores da mesma área e é citado de acordo com a sua importância. Um artigo menos citado nos estágios iniciais pode não chamar a atenção de outros pesquisadores e não ser citado. Isto nos dá uma realimentação positiva. Assim, o artigo mais citado inicialmente chama a atenção de mais pessoas e então é citado mais vezes, o que atrai mais pesquisadores que os citam mais vezes e assim vai... Portanto, o índice de citação cresce mais rapidamente para artigos importantes.

2.2.2. Geometria Fractal

O termo **“fractal”** (deriva do adjetivo *fractus*, do verbo *frangere,* que significa quebrar) foi introduzido na década de 70 pelo matemático Benoit Mandelbrot para designar objetos e estruturas complexas dotadas de propriedade de auto-similaridade (que nunca perdem a sua estrutura qualquer que seja a distância de visão) e dimensões fracionárias. (Addison, 1997).

Mandelbrot percebeu que é quase impossível descrever a natureza usando apenas a geometria Euclidiana, ou seja, em termos de linhas retas, círculos, cubos, etc. Portanto, ele criou a geometria fractal para descrever e analisar a complexidade das formas irregulares do mundo natural que nos cerca, como por exemplo, as árvores, os raios, a costa de um país, etc.

O contorno de nuvens, as rachaduras em uma parede, o contorno de montanhas, a folha de uma samambaia, entre outros, são também exemplos de fractais naturais. Dizemos que os três primeiros exemplos possuem **auto-similaridade estatística** (possuem o mesmo grau de irregularidade) e, no caso da folha de samambaia esta possui **auto-similaridade exata** (cada pedaço da folha é uma mini-cópia do todo).

Os **fractais regulares** são objetos que apresentam **auto-similaridade exata**.

Fractais ao acaso ou aleatórios possuem **auto-similaridade estatística**, e diferentemente dos fractais regulares, contêm na sua formação um elemento estatístico ou ao acaso, ou seja, cada pequena parte do fractal tem as mesmas propriedades estatísticas do todo.

Embora muitos estudos tenham sido realizados com fractais, não existe qualquer lei geral da Física que possa explicar o surgimento da geometria fractal na natureza.

2.2.3. Ruído 1/f

O ruído 1/f é um tipo de ruído cujo espectro de freqüência segue uma lei de potência que o relaciona inversamente com a sua freqüência f. A potência da componente de freqüência é maior para as freqüências menores, sendo inversamente proporcional à freqüência, é por isso que o chamamos de comportamento 1/f, embora o correto seria tratá-lo como sinal e não como comportamento.

Hurst (1951) passou parte de sua vida estudando o nível de água do rio Nilo, em várias escalas de tempo, (de minutos até anos) e observou que esta série temporal pode ser vista graficamente como uma superposição de todas as variações do nível do rio, ou seja, com uma superposição de sinais periódicos de todas as freqüências.

Existem casos em que o gráfico de freqüência não pode ser distribuído como sinal 1/f, mas como $1/f^{\alpha}$, onde α é um expoente com valor entre 0 e 2, e f é a freqüência, mas continuam sendo denominados como ruído 1/f.

Este tipo de ruído 1/f pode ser encontrado em uma grande variedade de sistemas complexos:

- variação de preço da bolsa de valores;
- tráfego de uma rodovia;
- temperatura média global;
- luminosidade das estrelas, etc.

2.2.4. Lei de Zipf

O professor George Kingsley Zipf (1949) fez várias observações para algumas regularidades simples de sistemas na área de ciências humanas. Trabalhando com número de habitantes de cidades, ele tomou a cidade com maior número de habitantes e atribuiu o rank r = 1, à segunda maior cidade o rank r = 2 e assim sucessivamente. Traçando o gráfico log do rank versus log do número de habitantes, obteve uma linha reta com inclinação próxima a 1.

Da mesma forma Zipf trabalhou com a freqüência de aparecimento de determinadas palavras em um livro. À palavra com maior freqüência foi atribuído o rank = 1, e assim sucessivamente de forma decrescente. Da mesma forma, através de um gráfico log do rank versus log da freqüência de cada palavra, obteve como resultado uma reta com inclinação próxima a 1.

Portanto, a lei de Zipf consiste em que existem sistemas nos quais determinadas variáveis são inversamente proporcionais à sua classificação, ou seja, $f(r) \sim r^{-\alpha}$ onde r é o rank e f é a freqüência de determinada ocorrência, com $\alpha \sim 1$.

Capítulo 3

Distribuições Estatísticas

Neste capítulo, mostraremos algumas distribuições estatísticas que são de grande interesse para a Física, e em particular, estão relacionadas ao estudo dos sistemas complexos.

3.1. Distribuição Normal

A distribuição normal surgiu no século XVIII ligada ao estudo de erros de medições repetidas de uma mesma quantidade. As suas propriedades matemáticas foram estudadas por De Moivre (1733) com o objetivo de aproximá-la à distribuição

binomial; por Laplace (1781), que usou a curva normal para descrever a distribuição dos erros; e, Gauss (1809) usou-a para analisar dados de astronomia, sendo por este fato, esta distribuição ser conhecida por *distribuição de Gauss ou distribuição Gaussiana*. Muitas vezes, esta distribuição é também apelidada de *curva em forma de sino*, como mostra a figura abaixo.

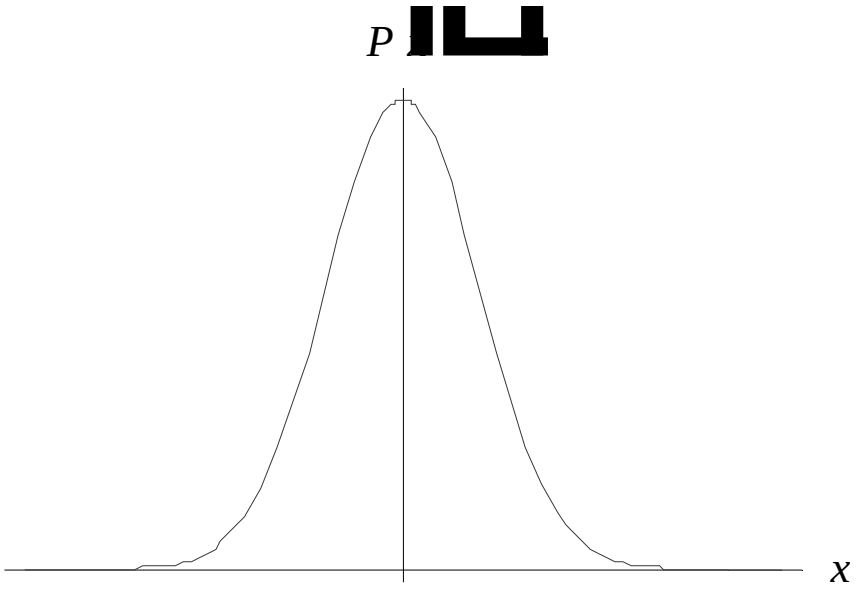

Figura 1. Distribuição Normal

Depois que Gauss se utilizou desta distribuição de probabilidades, os pesquisadores verificaram que quase a totalidade dos fenômenos que ocorrem na natureza, na Física, na Biologia e nas Ciências Sociais, segue esta distribuição, daí o termo “normal”. A *normalidade* é de fato muito importante na inferência estatística.

As principais razões da sua importância prendem-se ao fato de muitas variáveis físicas, biométricas, econômicas ou sociais, serem aproximadamente normais, e mesmo variáveis não normais poderem ser transformadas nesta, ou ainda, neste caso a parte central ser razoavelmente bem aproximada por uma normal. Esta distribuição é estável pelo Teorema do Limite Central.

A distribuição normal é dada por P(x):

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < +\infty, -\infty < \mu < +\infty, 0 < \sigma < +\infty \qquad (3.1)$$

onde P(x) é a densidade de probabilidade da variável aleatória contínua x, μ é a média aritmética e σ o desvio padrão populacional (ou equivalentemente a variância populacional σ^2). As estimativas de μ e σ são dadas por:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{e} \qquad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2} \qquad (3.2)$$

A média refere-se ao centro da distribuição e o desvio padrão ao espalhamento de curva. A distribuição normal é simétrica. O importante é que, a curva é afetada pelos valores numéricos de μ e σ, isto é mostrado no diagrama abaixo.

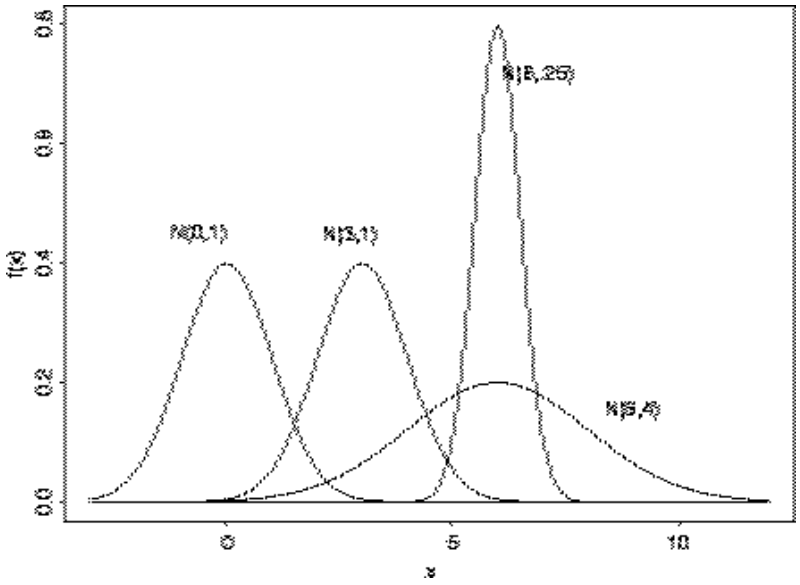

Figura 2. Média e Desvio Padrão na Distribuição Normal

A área sob a curva normal (na verdade abaixo de qualquer função de densidade de probabilidade) é 1. Então, para quaisquer dois valores específicos podemos determinar a proporção de área sob a curva entre esses dois valores.

Para a distribuição Normal, as proporções de valores caindo dentro de um, dois, ou três desvios padrão da média são:

Variação	Proporção
$\mu \pm 1\sigma$	68.3%
$\mu \pm 2\sigma$	95.5%
$\mu \pm 3\sigma$	99.7%

Na prática, desejamos calcular probabilidades para diferentes valores de μ e σ. Para isso, a variável x cuja distribuição é P(x) é transformada numa forma padronizada Z, que denominamos **distribuição normal padrão**, pois tal distribuição é tabelada. Assim, se considerarmos

$$Z = \frac{x - \mu}{\sigma} \tag{3.3}$$

então P(Z) é dada por:

$$P(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}} \tag{3.4}$$

A maioria das tabelas da distribuição normal são fornecidas somente para valores positivos da variável. Os valores negativos são encontrados pela simetria da distribuição em torno do zero.

Um dos teoremas mais importantes sobre as distribuições estatísticas é o **Teorema do Limite Central**, onde temos que se x_1, x_2, x_3, ..., x_n são valores independentes de uma variável x, com valor médio e desvio padrão finitos (condição necessária para todos os sistemas naturais), então a distribuição de S, onde $S = \sum_{i=1}^{n} x_i$ tende à Distribuição Normal para grandes valores de n, isto é, as distribuições estatísticas convergem para a normal quando o número de elementos amostrais n, tende à infinito.

3.2. Distribuição Log-normal

Uma distribuição freqüentemente associada à distribuição com caudas longas é a distribuição Log-normal, definida por

$$P(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2} \tag{3.5}$$

As estimativas de μ e σ são dadas por:

$$\mu_{\log} = \frac{1}{n}\sum_{i=1}^{n} \log(x_i) \qquad \text{e} \qquad \sigma_{\log} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\log x_i - \mu_{\log}\right)^2} \tag{3.6}$$

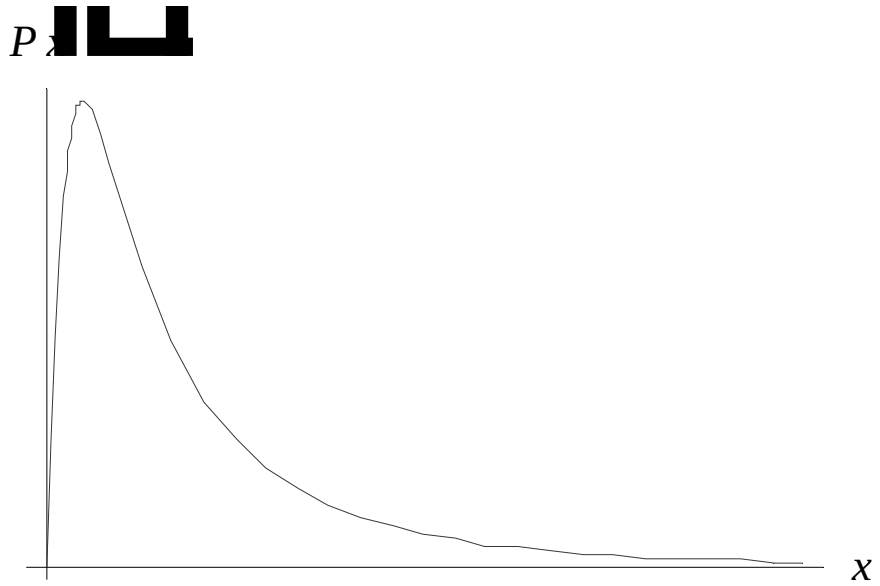

Figura 3. Distribuição Log-normal

Shockley (1957) propôs o seguinte mecanismo para explicar porque caudas longas existem em uma distribuição que exija um bom resultado para diversos empreendimentos, onde o fracasso de um provocaria o fracasso do projeto. Ele usou a publicação de documentos técnicos como exemplo. Considerou como importantes algumas habilidades abaixo mencionadas.

1. Habilidade de obter um bom tema;
2. Habilidade para trabalhar nele;
3. Habilidade para reconhecer soluções que valham a pena;
4. Habilidade de tomar decisões de quando parar ou obter resultados;
5. Habilidade de redigir adequadamente;
6. Habilidade de aproveitar-se construtivamente das críticas;
7. Determinação para apresentar o trabalho em jornais;
8. Disposição para agüentar julgamentos de oposição.

Ele relata assim, que a probabilidade de um pesquisador produzir um trabalho de sucesso em tempo determinado seria o produto de um conjunto de probabilidades que provocaria o sucesso do empreendimento, ou seja, possuindo cada uma das habilidades da relação anterior o sucesso seria evidente.

$$P = p_1.p_2.p_3.p_4.p_5.p_6.p_7.p_8 \qquad (3.7)$$

O aspecto log-normal torna-se aparente ao considerarmos logaritmos na equação acima,

$$\log p = \log p_1 + \log p_2 + \ldots + \log p_8 \qquad (3.8)$$

Desde que log p seja a soma de um conjunto de variáveis, cada qual com sua própria função de distribuição, o teorema do limite central é aplicável de modo que a distribuição da função log p poderia ser Gaussiana.

Durante os últimos 90 anos, a distribuição log-normal tem sido observada em muitos sistemas, incluindo a Ecologia, a Medicina, o Meio Ambiente, a Lingüística, entre outros, e várias explicações foram dadas aos mesmos, além do proposto por nós nas equações 3.7 e 3.8. Veja alguns exemplos:

- **Lingüística –** O número de letras por palavra e o número de palavras por sentença seguem uma distribuição log-normal (Herdan, 1958; Williams, 1940).

- **Tecnologia de Alimentos –** Várias aplicações da distribuição log-normal são relatadas para caracterização de estruturas em tecnologia de alimentos e engenharia no processamento de alimentos (Reinders et. al., 2002).

- **Ecologia –** Na maioria das comunidades de plantas e animais, a abundância de espécies segue a distribuição log-normal (truncada) (Sugihara, 1980; Magurran, 1988).

- **Ciências Sociais e Econômicas –** Exemplos de distribuição log-normal nas Ciências Sociais e Econômicas incluem idade de casamento, tamanho de fazendas e rendimentos. (Preston, 1981).

- **Fisiologia das plantas –** Recentemente, evidências convincentes foram apresentadas vindas da fisiologia de plantas, indicando que existe uma

distribuição log-normal para a permeabilidade de água nas folhas (Baur, 1997).

- **Fitomedicina e Microbiologia –** Exemplos vindos da Microbiologia e da Fitomedicina incluem a distribuição de sensibilidade a fungicidas em populações e distribuição do tamanho de populações (Romero e Sutton, 1997).

- **Ciência Atmosférica e Aerobiologia –** A qualidade do ar está contida nos microorganismos, que eram muito maiores e menos variáveis no ar de Marseille do que o de uma ilha (Di Giorgio et al.,1996). A atmosfera é a principal parte do sistema de manutenção da vida, e muitas propriedades físicas e químicas da atmosfera seguem a lei da distribuição log-normal (Limpert et al., 2000).

- **Meio Ambiente –** A distribuição de partículas, produtos químicos e microorganismos no meio ambiente freqüentemente seguem a distribuição log-normal (Biondini, 1976).

- **Medicina Humana –** Uma variedade de exemplos vindos da medicina seguem a distribuição log-normal. Períodos latentes (tempo da infecção até o primeiro sintoma) de doenças infecciosas são freqüentemente descritas pela distribuição log-normal. (Sartwell,1950, 1952, 1966; Kondo, 1977). Tempo de sobrevivência após o diagnóstico de um câncer (Boag, 1949).

- **Geologia e Mineração –** Na crosta-terrestre, a concentração de elementos e sua radioatividade usualmente seguem a distribuição log-normal (Razumovsky, 1940; Ahrens, 1954; Malanca et al., 1996).

3.3. Distribuição Exponencial

Dizemos que uma variável aleatória t tem distribuição exponencial de parâmetro λ se a sua função densidade de probabilidade for dada por:

$$f(t)=\begin{cases}0 & se \quad t<0\\ \lambda e^{-\lambda t} & se \quad t\geq 0\end{cases} \tag{3.9}$$

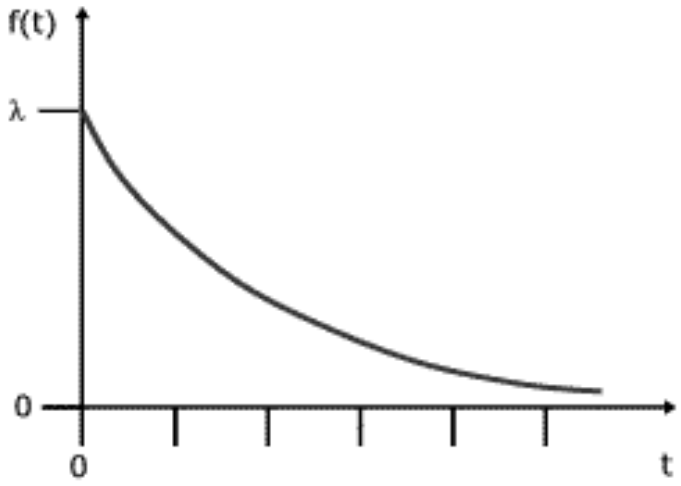

Figura 4. Distribuição Exponencial

As estimativas de μ e σ são dadas por:

$$\mu_{\exp}=\frac{1}{\lambda} \qquad \text{e} \qquad \sigma_{\exp}=\frac{1}{\lambda} \tag{3.10}$$

Esta distribuição pode descrever inúmeros fenômenos físicos, como o tempo t para o decaimento de um núcleo radioativo, ou o tempo x para um componente falhar, ou ainda a distância z no curso de um fóton na atmosfera antes de sofrer uma colisão com uma molécula de água.

3.4. Distribuição de Lei de Potência

A distribuição de Lei de Potência (Lévy, 1937; Pareto, 1896) foi primeiramente notada por Pareto na renda pessoal (Pareto, 1896), e posteriormente por outros em muitos sistemas complexos físicos (Solomon, 1993; Chabaud et. al., 1994), biológicos (Peng, 1993; Bassingthwaighte, 1994), econômicos (Ott, 1990; Mandelbrot, 1963) e educacionais (Gupta et. al., 2003).

Pareto considerou que esta distribuição é devida à realimentação positiva, isto é, por exemplo, o rico pode mais eficientemente elevar sua riqueza que um indivíduo de classe média.

Tsallis (1999) obteve a distribuição de Lei de Potência através da termodinâmica não-extensiva em que eles incorporaram a interação de longa escala e a memória microscópica de longa escala. Tsallis (1999) propôs uma definição generalizada de entropia como segue:

$$S = k\frac{\sum_i p_i^q}{q-1} \quad \left(\sum_i p_i = 1; q \in \Re\right) \tag{3.11}$$

onde k é uma constante positiva. Mais adiante (Capítulo 4), veremos que Tsallis e Albuquerque deduziram a fórmula da densidade de publicação baseados nesta definição de entropia.

A distribuição Lei de Potência é dada por

$$P(S) = C.S^{-\alpha} \tag{3.12}$$

onde P(S) é a probabilidade de ocorrência do evento S; α é o expoente da distribuição e C é uma constante. Então, temos que

$$\log P(S) = -\alpha \log (S) + k \text{ , onde k é constante.} \tag{3.13}$$

Assim, podemos dizer que os gráficos do tipo log x log que apresentam como resultado uma linha reta, descrevem a distribuição de Lei de Potência com α sendo a inclinação desta reta.

Um exemplo desta distribuição é a magnitude dos terremotos. Os terremotos distribuem-se em função da energia liberada, de acordo com a lei de potência (Geller, 1997).

Se N(E) é o número anual de terremotos em uma determinada região, onde E é a energia, então temos que $N(E) = E^{-b}$, com a constante b ~ 2. Esta relação é chamada lei de Gutenberg-Richter (relação estatística para as observações). Portanto, isto não especifica quando um terremoto de determinada magnitude irá ocorrer, mas somente a distribuição da sua magnitude. Aplicando o logaritmo em ambos os lados da equação obtemos uma linha reta. A constante b nos dá o declive desta linha.

Curiosamente, a distribuição da extinção de espécies na Terra (Vines, 1999) é idêntica à lei de potência da distribuição de terremotos (dobrando o tamanho da extinção – medida pelo número de famílias extintas – ela torna-se quatro vezes mais difícil de ocorrer).

Outros exemplos que seguem de Lei de Potência são:

- as guerras (Lévy, 1983);
- os incêndios em florestas (Malamud et. al., 1998);
- a citação de artigos (Redner, 1998);
- a expansão de tijolos devido à umidade (Wilson, 2003), e muitos outros.

3.5. Distribuição de Lei de Potência Gradualmente Truncada

A distribuição de Lei de Potência possui desvio padrão infinito, embora os sistemas físicos reais apresentem desvio padrão finito. Ao aplicarmos então a distribuição de Lévy, Pareto ou ainda em alguns casos a Log-normal, aos sistemas físicos reais, precisamos truncar as distribuições após um determinado valor, a fim de evitar um desvio padrão infinito.

Pareto propôs a distribuição de lei de potência baseado na realimentação positiva. Gupta e Campanha (1999), consideram que a validade da Lei de Potência tem um limite devido à capacidade física do sistema, e portanto, a realimentação positiva também deveria cessar após um certo valor crítico de alguma variável. Como em sistemas complexos temos várias interações e um grande número de componentes interagindo de maneiras diferentes, esperamos que o truncamento desta realimentação positiva seja gradual após um ponto crítico.

A distribuição proposta por Gupta e Campanha (1999, 2000), denominada *Distribuição de Lei de Potência Gradualmente Truncada* é dada por,

$$P_{GT}(x) = P(x) f(x) \tag{3.14}$$

onde P(x) é a distribuição de Lei de Potência:

$$P(x) = \frac{c_1}{c_2 + \left(|x - x_m|\right)^{1+\alpha}} \tag{3.15}$$

P(x) é a probabilidade de x, x_m é o valor de x onde a probabilidade é máxima, c_1 e c_2 são constantes:

$$c_1 = c_2 P(x_m) \tag{3.16}$$

e c_2 pode ser obtido através da condição de renormalização.

E, $f(x)$ é dada por

$$f(x)=\begin{cases}1 & se \quad |x| \le x_c \\ e^{-\left(\frac{|x-x_c|}{k}\right)^{\beta}} & se \quad |x| > x_c\end{cases} \tag{3.17}$$

onde α é o índice de lei de potência, x_c é o ponto crítico onde começa o truncamento gradual devido ao limite físico do sistema (geralmente é muito maior do que x_m), e k é uma constante de truncamento gradual. Para valores menores de k, o truncamento será mais rápido. Nós escolhemos $\beta = 2 - \alpha$, assim como mostrado por Gupta e Campanha (2000), para que essa distribuição se aproxime da Distribuição Normal para grandes escalas, o que é condição essencial para qualquer distribuição pelo Teorema do Limite Central.

Esta distribuição tem desvio padrão finito e também variância finita, no limite obtém-se uma Distribuição Normal como exigido pelo Teorema do Limite Central (Gupta e Campanha, 2000) e, além disso, também obedece a Lei de Potência em sua parte central e decaimento exponencial nos valores extremos de x. Podemos considerar esta distribuição equivalente à distribuição de Tsallis, obtida pela termodinâmica para sistemas complexos. A Distribuição Gradualmente Truncada torna mais simples a extração de informações úteis que descrevem o sistema real.

3.6. Gráfico de Zipf

Geralmente, temos muito poucos dados nos extremos das distribuições estatísticas dos sistemas reais que estudamos, isto dificulta saber que tipo de distribuição ajusta-se a estes valores, pois qualquer distribuição parece cabível dentro do limite de erro. Neste caso, usaremos a técnica denominada de gráfico de Zipf, para a distribuição nesta faixa.

Suponhamos x_1, x_2, ..., x_n como o conjunto de N observações de uma variável x, ordenadas de forma decrescente, isto é, o índice i nos dá a posição da

observação x_i no ranking. Se a probabilidade é dada por P(x), então o índice i de x pode ser calculado por

$$i = N\int_{x_i}^{\infty} P(x)dx \tag{3.18}$$

onde $\int_{x_i}^{\infty} P(x)$ nos dá a probabilidade de observações que têm valores acima de x_i.

A multiplicação por N nos dá o número de observações igual ou acima de x_i, isto é, a classificação das observações de x_i. Desta forma, podemos saber se a distribuição assumida é certa ou não. Assim, o gráfico na escala logarítmica aumenta significantemente para os altos valores da distribuição, ficando mais fácil comparar a teoria com os resultados empíricos nesta faixa. Por exemplo, na Figura 5, numa amostra de 1000 dados, os 10 dados mais altos ocupam 33% do espaço numa dupla escala logarítmica ao invés de 1% em um gráfico normal, como mostrado na Figura 6. Esta técnica tira a flutuação nesta faixa e facilita uma análise quantitativa.

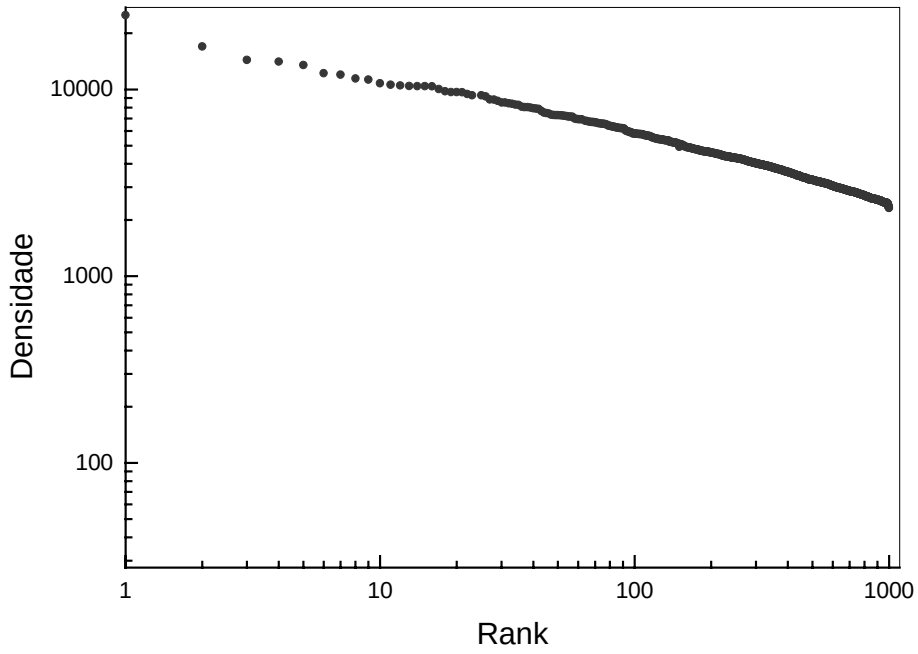

Figura 5. Conjunto de 1000 observações de um determinado fenômeno real, numa dupla escala logarítmica.

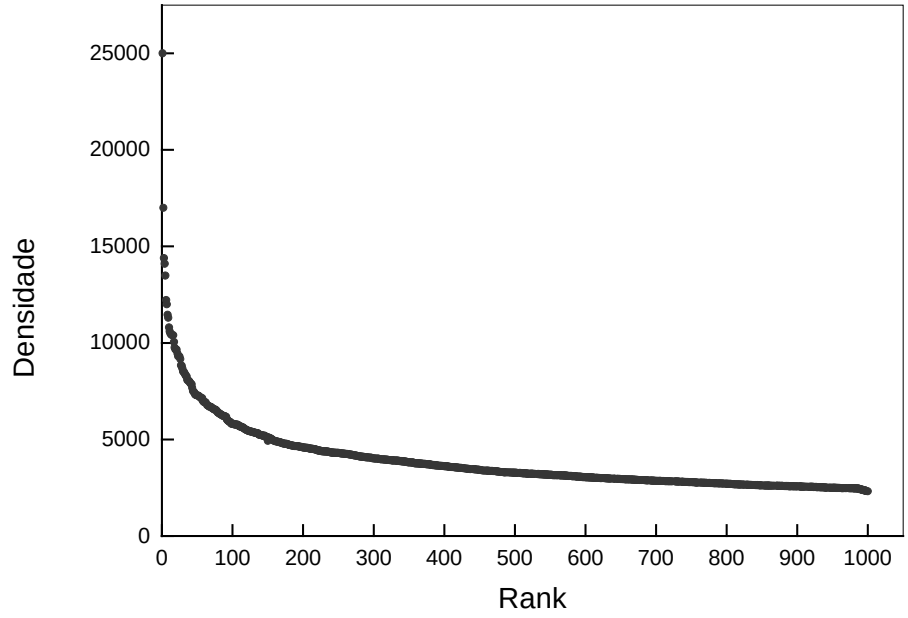

Figura 6. Conjunto de 1000 observações do mesmo fenômeno real mostrado na Figura 5, numa escala normal.

Capítulo 4

A Distribuição do Índice de Citação de Publicações Científicas

4.1. Introdução

Sabemos que a publicação científica é o meio mais importante de comunicação entre as diversas áreas da Ciência e que, o mérito de um cientista ou de um artigo científico é normalmente considerado através do número de vezes em que ele ou o seu artigo são citados em outros trabalhos científicos. Embora não seja uma medida exata da importância de um artigo ou de um cientista, esta pode ser considerada uma boa medida (Redner, 1998).

Em 1957, num estudo baseado no registro do pessoal da pesquisa científica do Brookhaven National Laboratory, Schockley (1957) afirmou que a taxa de publicação científica é descrita por uma distribuição log-normal.

Laherrere and Sornette (1998) apresentaram evidência numérica, baseada nos dados de 1120 físicos mais citados desde Janeiro de 1981 a Junho de 1997, que a distribuição acumulada de citação de autores individuais tem a forma Exponencial Estendida,

$$N_c(x) = k.\exp\left(-\frac{x}{x_0}\right)^{\beta}$$, onde k = constante. (4.1)

A Exponencial Estendida é caracterizada pelo valor de c < 1. No caso de c = 1, ela se transforma numa Distribuição Exponencial. No caso dos físicos mais citados, eles obtiveram β = 0,3.

Usando a técnica do gráfico de Zipf, Redner (1998) recentemente mostrou que a distribuição de citação dos artigos científicos mais citados é descrita pela Lei de Potência

$$N(x) \sim x^{-\alpha}$$ com $\alpha \cong 3.0$. (4.2)

Tsallis e Albuquerque (2000) mostraram que a distribuição de citação de artigos científicos é descrita pela Estatística de Tsallis.

Simkin e Roychowdhury (2003), propõem um modelo de citações considerando que um cientista, ao escrever um trabalho, copia uma fração da referência de outros trabalhos sem lê-los. Assim, um artigo citado inicialmente tem uma grande chance de citação em outros artigos. Em outro trabalho, Bagrow et al. (2004) mediu a fama pelo número de acessos no site Google, e o mérito pelo número de artigos colocados num arquivo eletrônico. Bagrow observou que a fama é proporcional ao mérito de um cientista, e que a distribuição de probabilidade da fama cai exponencialmente. Num trabalho recente, Redner (2005) estudou a estatística

das citações para trabalhos publicados em 110 anos de Physical Review. Ele mostrou como os artigos são citados, além disso, descreveu a fascinante história de citação de alguns artigos.

O número de publicações científicas e o de citações de um pesquisador evidentemente não são os mesmos, pois possuem mecanismos diferentes. O primeiro representa a quantidade de trabalho, enquanto que o segundo representa a qualidade e originalidade dos trabalhos.

Novamente, os índices de citação de uma publicação científica e de um cientista também não são os mesmos. O índice de citação de uma publicação científica simplesmente depende da qualidade e originalidade do trabalho e, a importância dele para outros trabalhos. O índice de um cientista é a combinação de qualidade e quantidade de seus trabalhos. Assim, o número de publicações de um cientista, o índice de citação de um cientista e o índice de citação de uma publicação científica têm mecanismos diferentes e assim podem ter distribuições diferentes. Como todos os três parâmetros são importantes na carreira de um cientista, iremos discuti-los separadamente.

O número de trabalhos científicos publicados por um cientista depende do produto de muitas habilidades, como:

(i) escolha do tema apropriado;
(ii) profundidade no tratamento deste problema;
(iii) escolha de um veículo apropriado para a divulgação dos resultados;
(iv) habilidade e objetividade de redação, etc.

E, então como apontado por Shockley (1957), a distribuição deveria ser log-normal.

No presente capítulo, discutiremos a densidade de publicação científica em função da citação. No capítulo seguinte, trabalharemos com o índice de citação dos cientistas.

Em ambos trabalhos sobre citação de publicação científica (Redner, 1998; Tsallis e Albuquerque, 2000), o número de publicações versus citações é colocado em um gráfico, e os zeros são simplesmente ignorados.

Baseados na definição de entropia, dada pela equação 3.11 e da otimização da mesma com a condição de renormalização com $\sum_{x=1}^{\infty} p_x = 1$, e, $\left\langle x_q \right\rangle \equiv \frac{\sum_{x=1}^{\infty} x p_x^q}{\sum_{x=1}^{\infty} p_x^q} = const.$, temos a Estatística de Tsallis (2000) dada por:

$$N(x) = \frac{N_0}{\left[1 + (q-1)\lambda x\right]^{\frac{q}{q-1}}} \tag{4.3}$$

onde $N(x)$ no presente caso é a densidade de probabilidade, q e λ são parâmetros livres e N_0 é a constante de renormalização. Ainda podemos simplificar isto como:

$$N(x) = \frac{N_0}{\left[1 + c_1 x\right]^{(1+\alpha)}} \tag{4.4}$$

onde c_1 é uma constante e $(1+\alpha)$ é o índice da Lei de Potência. Para grandes valores de x, a distribuição fica da seguinte forma:

$$N(x) \approx c x^{-(1+\alpha)} \tag{4.5}$$

isto é, uma Lei de Potência. Neste caso, $\log N(x)$ versus $\log x$ é uma linha reta para grandes valores de x.

A Lei de Potência não pode continuar para sempre em sistemas reais. Esta tem que ser truncada em algum lugar para impedir a variância e/ou outros momentos infinitos. No caso das publicações científicas, o campo torna-se saturado ou quase inteiramente investigado depois de um certo tempo entre 20 e 100 anos dependendo do campo. Assim, os pesquisadores da área também começam a

reduzir depois deste período e também as citações. No caso de um cientista, além da saturação do campo, existem as limitações humanas de produzir muitos trabalhos científicos importantes.

Geralmente, a densidade de publicação é muito baixa para artigos altamente citados, portanto é interessante construir um gráfico de Zipf (Galambos, 1978), em que o número de citações do n-ésimo artigo mais citado dentre um conjunto de M artigos é colocado no gráfico versus o rank n. Sendo assim, o gráfico de Zipf está relacionado com a probabilidade acumulada da cauda longa de x da distribuição de citação. Este gráfico satisfaz portanto a determinação da cauda longa de x da distribuição de citação, além de suavizar a flutuação da cauda dos mais citados, facilitando assim a análise quantitativa.

Dado um conjunto de M artigos e o correspondente número de citações para cada um destes artigos no rank ordenado $Y_1 \geq Y_2 \geq Y_3 \ ... \geq Y_n \geq ... \ Y_M$, então o número de citações do n-ésimo artigo mais citado Y_n pode ser estimado pelo critério (Gupta e Campanha, 2000):

$$\int_{Y_n}^{\infty} N(x)dx = \int_{Y_n}^{\infty} M.P(x)dx = n \qquad (4.6)$$

Isto especifica que existem n artigos dentro de um conjunto de M que são citados no mínimo Y_n vezes. Desta dependência de Y_n sobre n num gráfico de Zipf, pode-se testar se isto está de acordo com a distribuição suposta para N(x).

4.2. Análise de dados

Neste caso, estamos trabalhando com um dos maiores conjuntos de dados de publicação científica fornecido pelo ISI (Institute for Scientific Information), que contém a distribuição de citação de 783399 artigos (com 6716198 citações) publicadas em 1981 e citadas entre Janeiro de 1981 e Junho de 1997, em todos nos níveis de jornais (Dados do site http://physics.bu.edu/~redner/projects/citation/isi.html).

Em ambos os trabalhos de Redner (1998) e Tsallis (2000) sobre a citação de publicação científica, são feitos gráficos do número de publicações versus as citações e os zeros são simplesmente ignorados. Numa distribuição estatística, obtemos a densidade de publicação, isto é, o número de publicações num intervalo de unidade de citação e não o número de publicações. Isto faz diferença na distribuição dos artigos mais citados quando temos poucos artigos nesta escala. Além disso, os zeros não podem ser ignorados pois $\ln(0) = -\infty$.

Assim, é necessário fazer um gráfico da densidade de publicação versus a citação para ter a correta visão dos mecanismos da citação. Fizemos o gráfico da densidade de publicação, isto é, o número de publicações por citação versus citação.

A densidade de publicação é dada por:

$$N(x) = \left[\frac{\Delta N}{\Delta x}\right]_x \tag{4.7}$$

onde ΔN é o número de publicações que têm citações entre $\left(x - \frac{\Delta x}{2}\right)$ e $\left(x + \frac{\Delta x}{2}\right)$.

Para índices de citação mais baixos, tomamos $\Delta x = 1$, onde nós temos um grande número de publicações, ao passo que para grande índices de citação, aumentamos gradualmente Δx para ter somente valores diferentes de zero de densidade de publicação para citação.

Observamos que para quase oito graus de magnitude (10^{-4} a 10^{4}) da densidade de publicação, a distribuição é dada através da Estatística de Tsallis como mostrada na Figura 7, com N_0 = 4,66x10^4, c_1 = 0,0583 e 1 + α = 3,1. Os valores de N_0 e α são escolhidos através da inclinação da reta que melhor se encaixa para $x \geq 100$. c_1 é escolhido para ter um melhor encaixe nos estágios iniciais.

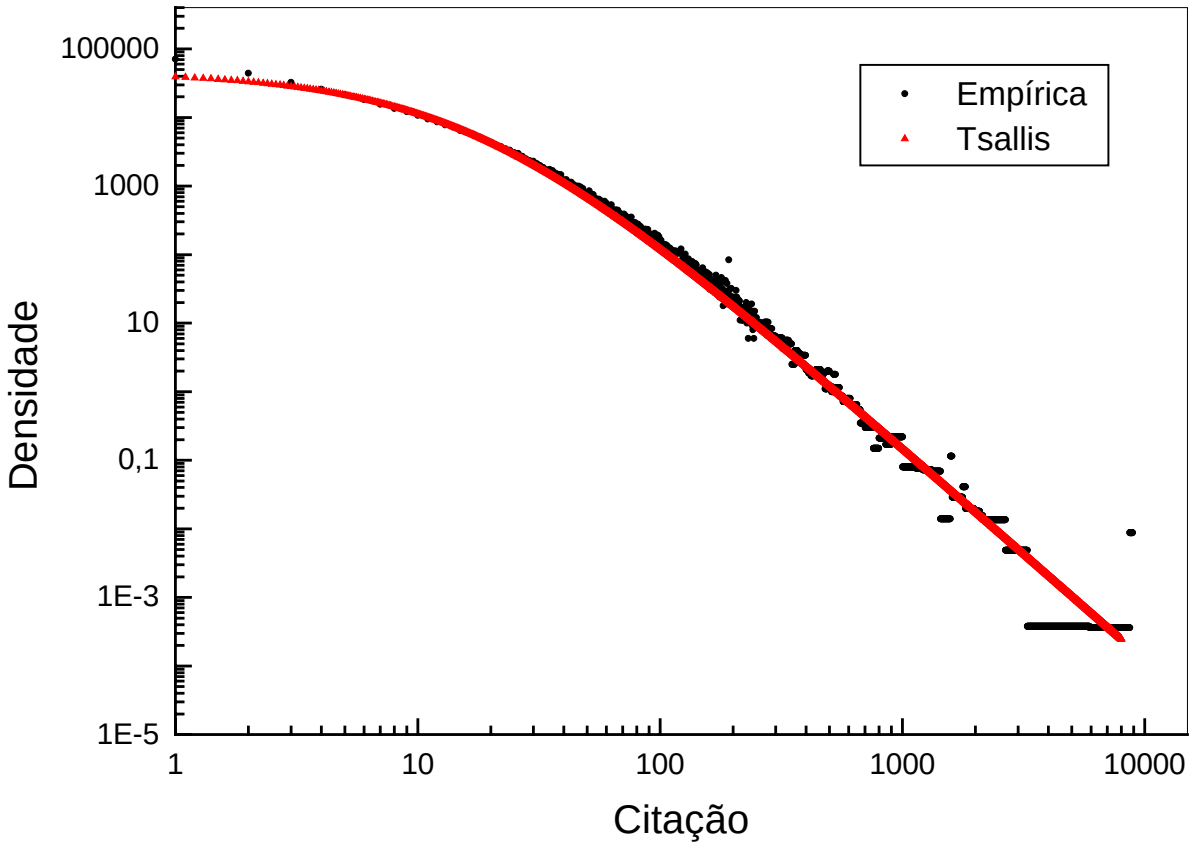

Figura 7. Densidade de publicação (publicação por citação) versus a distribuição de citação para 783339 artigos nos dados da ISI numa dupla escala logarítmica.

Portanto, para confirmar a suposta distribuição, fizemos o gráfico de Zipf na Figura 8, com os mesmos parâmetros. Este ajuste está excelente, mostrando que a distribuição é dada pela Estatística de Tsallis.

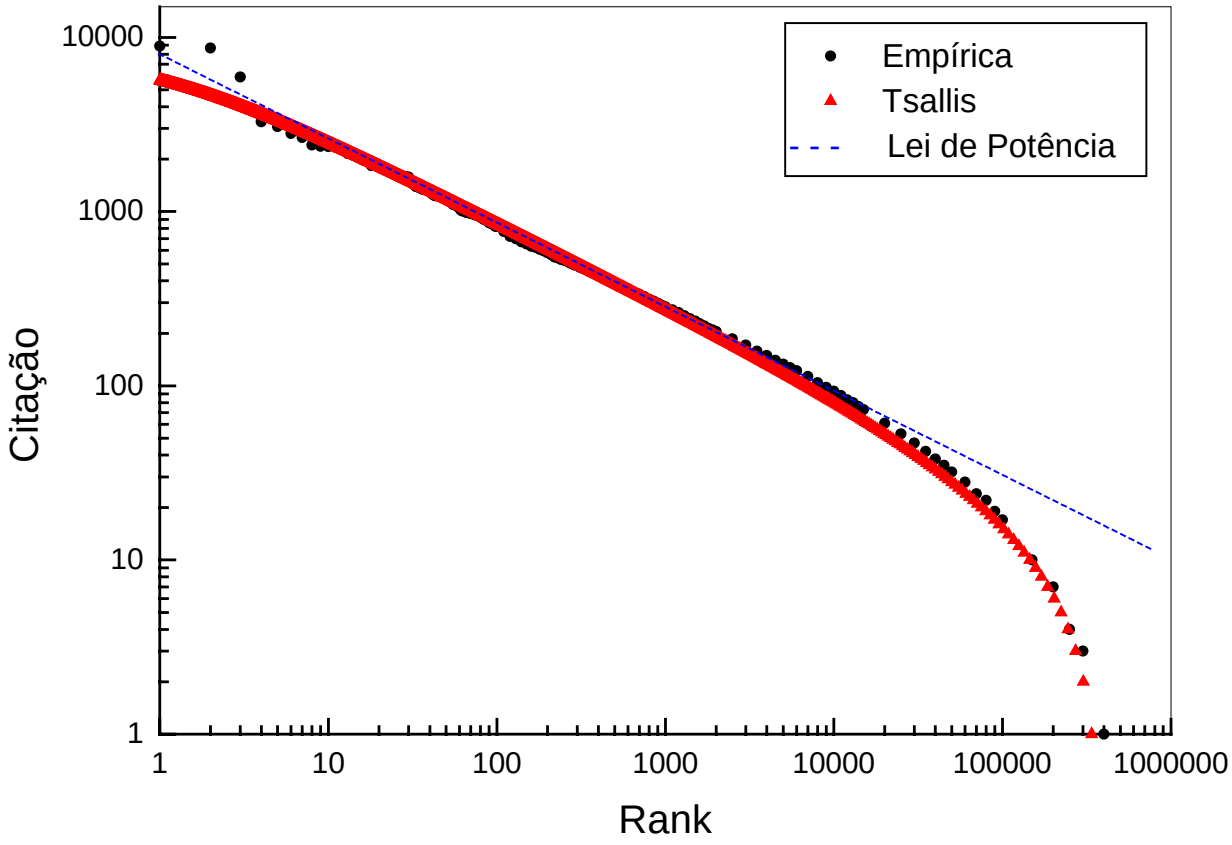

Figura 8. Gráfico de Zipf do número de citação do n-ésimo artigo no rank Y_n versus o rank n numa dupla escala logarítmica dos dados do ISI.

Na Figura 9, mostramos o gráfico log N(x) versus x^{β}. Escolhemos $\beta \cong 0.3$ que nos dá a melhor linha reta aproximada. Desenhamos uma linha reta para comparação. Isto mostra que a distribuição Exponencial Estendida não é apropriada para o presente caso.

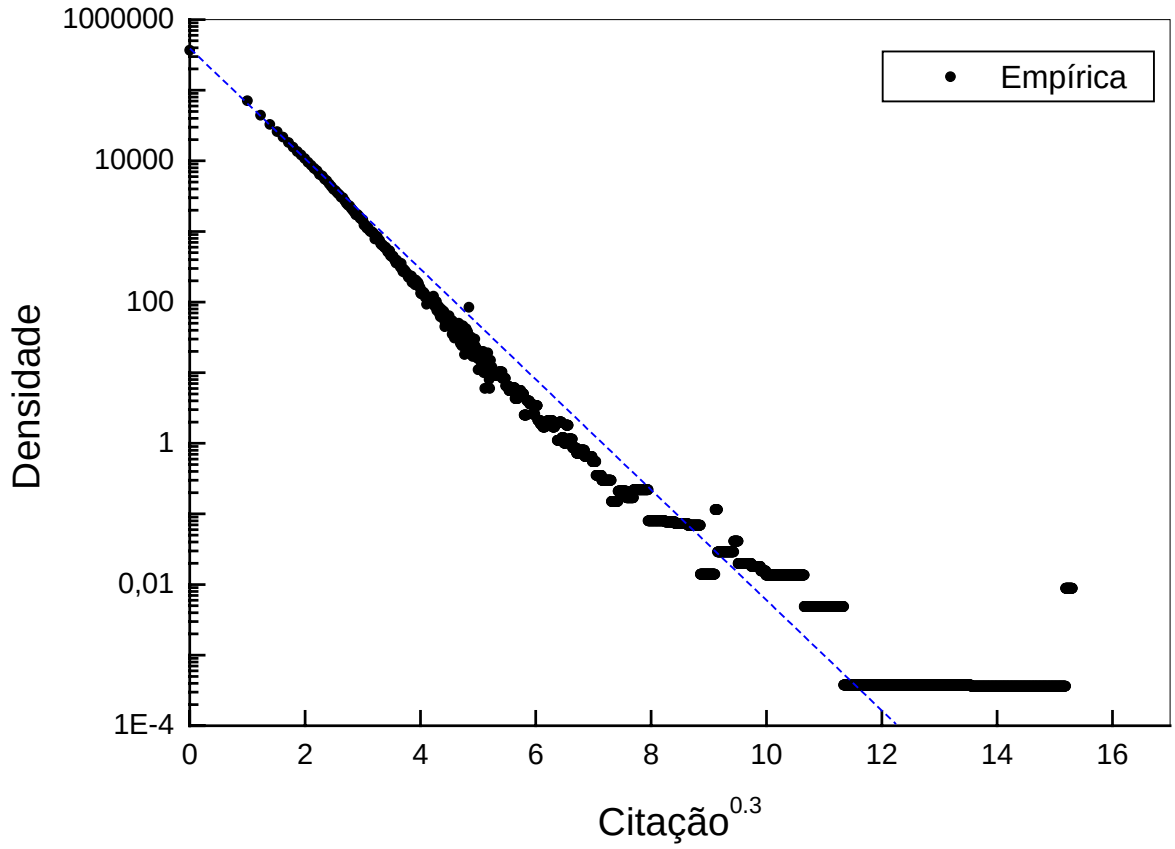

Figura 9. Logaritmo da densidade de publicação, versus (índice de citação)$^{0.3}$. O melhor ajuste em linha reta é desenhado para comparação.

Capítulo 5

A Distribuição do Índice de Citação dos Cientistas

5.1. Introdução

Como muitos acadêmicos são obrigados a documentar suas citações para consideração do seu mérito, é interessante entender a distribuição de citação e os mecanismos por trás disto.

No capítulo anterior, mostramos que a citação de um artigo científico é dada através da Estatística de Tsallis, o que dá a Lei de Potência para grandes valores de

citação. Neste capítulo, discutiremos o índice de citação individual dos cientistas, que é um importante parâmetro para a discussão da importância do seu trabalho, promoção e reputação.

O índice de citação de um cientista é a soma das citações de seus artigos. Pelo Teorema do Limite Central, isto possivelmente conduz para a Distribuição Normal, se o número de artigos de um cientista é muito grande e a citação de cada artigo é distribuída arbitrariamente. Contudo o número de artigos são limitados e se um importante cientista está tendo um alto índice de citação de quase todos os seus artigos, é por causa da sua boa e estabilizada reputação, e, o mérito do trabalho.

5.2. O modelo

Gupta e Campanha (2000), mostraram que através da Distribuição de Lei de Potência Gradualmente Truncada depois de certo valor crítico, puderam explicar toda a distribuição inclusive para grandes escalas em sistemas complexos físicos e financeiros. Consideraram que a Distribuição de Lei de Potência é devida a realimentação positiva que cessa gradualmente depois de certa escala através da capacidade física limitada dos componentes do sistema ou pelo próprio sistema (Gupta e Campanha, 2000). Esta distribuição se aproxima da Distribuição Normal no limite, e é uma importante candidata para o índice de distribuição de citação, que é descrita por:

$$N(x) = cx^{-(1+\alpha)} f(x) \tag{5.1}$$

onde c é uma constante e $f(x)$ é dada por

$$f(x) = \begin{cases} 1 & se \quad |x| \le x_c \\ e^{-\left(\frac{|x-x_c|}{k}\right)^{\beta}} & se \quad |x| > x_c \end{cases} \tag{5.2}$$

onde x_c é o valor crítico do tamanho da escala onde a distribuição de probabilidade começa a desviar da distribuição de Lei de Potência pelas limitações físicas, k é a constante de truncamento. Igualando esta distribuição com a distribuição normal, temos que β está relacionado com α, ou seja,

$$\beta = 2 - \alpha \tag{5.3}$$

Agora, vamos considerar dois casos especiais:

Caso I: quando $x \leq x_c$ então

$$N(x) = cx^{-(1+\alpha)} \tag{5.4}$$

isto é, temos uma Distribuição de Lei de Potência.

Caso II: quando $x >> x_c$.

Neste caso, a variação é devida a $f(x)$ estar predominando comparada a Lei de Potência e assim,

$$N(x) \approx e^{-\left(\frac{|x|}{k}\right)^{\beta}} \tag{5.5}$$

ou $\log N(x)$ versus x^{β} é uma linha reta. Isto nos dá uma Distribuição Exponencial Estendida.

No caso da Distribuição de Lei de Potência pura, usando a equação 4.5 e 4.6 temos que:

$$Y_n = c_1 M \alpha n^{-\frac{1}{\alpha}} \tag{5.6}$$

ou

$$\log Y_n = \left(-\frac{1}{\alpha}\right)\log n + b \tag{5.7}$$

isto é, $\log Y_n$ versus $\log n$ nos dará uma linha reta e b é uma constante.

Para a distribuição Exponencial Estendida, isto é, usando a equação 5.5 temos:

$$Y_n^{\beta} = -a \ln n + b \tag{5.8}$$

onde a e b são constantes. Neste caso, Y_n^{β} versus $\ln n$ pode ser uma linha reta.

Em geral, o índice de citação de cientistas altamente citados é acessível. Portanto, é interessante construir um gráfico de Zipf, em que o número de citações do k-ésimo mais posicionado cientista fora de um conjunto de M cientistas é graficado versus o rank k. Então, o gráfico de Zipf está aproximadamente relacionado com a probabilidade acumulada da cauda longa de x na distribuição de citação.

Consideramos que a distribuição de lei de potência gradualmente truncada também é válida no caso do índice de citação como em outros casos. O fato de um cientista ser citado mais vezes facilita a ele conseguir mais ajuda financeira para seus projetos investigativos e estudantes melhores, os quais formam melhores e maiores grupos. Como os artigos são inicialmente citados em artigos do mesmo grupo, qualquer artigo vindo de fora deste grupo tem mais citações no estágio inicial. Um artigo mais citado chama a atenção para outro cientista da mesma área e é citado por outros para completar a introdução do problema. O cientista mais citado pode mais eficientemente nivelar seu índice de citação que o cientista médio, criando mais citação e obtendo um alto nível do índice de citação. Assim, o mecanismo de realimentação positiva aumenta a produção de qualquer parâmetro que esteja sendo analisado. Este efeito de realimentação decresce gradualmente depois de certo tempo devido à limitação física do sistema. No presente caso a

limitação vem do tempo disponível limitado e da capacidade humana de um cientista.

5.3. Análise de Dados

Analisamos o índice de citação (a) dos físicos e químicos brasileiros mais citados e (b) físicos e químicos mais citados em todo o mundo. Todos os físicos (químicos) incluindo físicos (químicos) brasileiros publicam seus trabalhos nos mesmos jornais e trabalham quase nos mesmos problemas devido à natureza básica dos assuntos. A Física, como qualquer outra ciência básica, é a mesma em todo o mundo. Os físicos (químicos) brasileiros formam um pequeno grupo dentro da comunidade física (química), então, o fator limitante é importante somente para uns poucos cientistas no cume do “ranking”, pois o índice de citação de somente estes cientistas é superior ao valor crítico (x_c) da distribuição de Lei de Potência Gradualmente Truncada. Para o resto dos cientistas mais citados, somente a lei de potência é importante, e assim o índice de lei de potência pode ser bem avaliado.

No caso dos cientistas internacionais mais citados, o fator limitante pode ser importante para quase todos, pois o índice de citação de quase todos os cientistas é superior ao valor crítico (x_c).

Na Figura 10, fizemos o gráfico o número de citações (Y_n) versus o rank (n) para os primeiros 205 físicos brasileiros no ano de 1999 (Folha de São Paulo, 1999, site: http://www.uol.com.br/fsp/especial/ranking). Observamos uma linha reta para altos valores de n como é esperado no caso I, o qual começa a desviar para valores menores de n (n < 20), graficamos a curva teórica considerando a distribuição de Lei de Potência Gradualmente Truncada com α = 1,53, x_c = 2000, k = 1000, β = 0.47 e c = $2x10^6$. O ajuste da curva teórica com os resultados empíricos é bom.

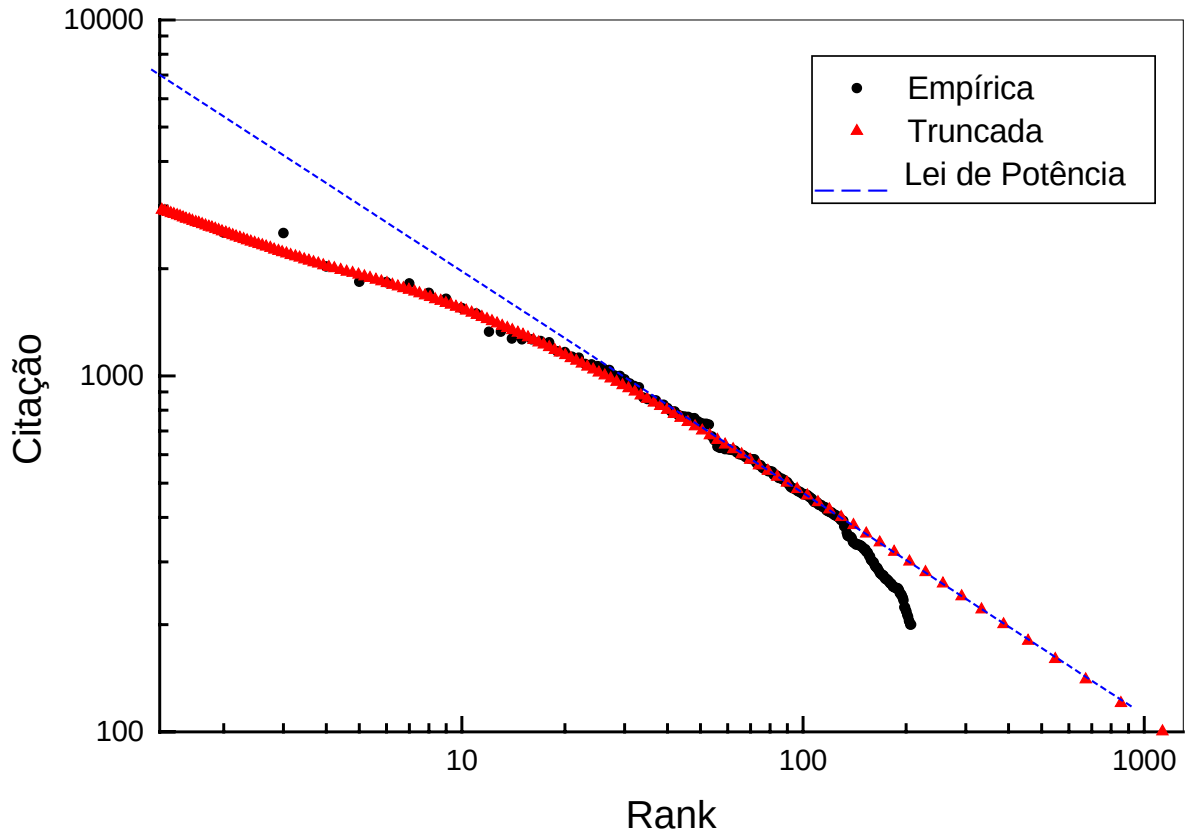

Figura 10. Gráfico log x log da dependência do n-ésimo rank Y_n como uma função do rank n, onde Y_n é o total de número de citações do n-ésimo físico brasileiro mais citado.

Na Figura 11 fizemos o gráfico do número de citações (Y_n) versus o rank (n) para 1.120 físicos mais citados no período de Janeiro de 1981 a Junho de 1997 (Dados do site http://physics.bu.edu/~redner/projects/citation/physics-by-person.html) com os mesmos valores de parâmetros usados na Figura 10 e comparamos isto com a curva teórica. Mudamos o valor da constante de $2x10^6$ para $1x10^9$ pois o número total de físicos neste caso é muito maior. Consideramos a citação de físicos brasileiros de aproximadamente 0.2% do total de citações, o que é razoável. Novamente observamos um bom ajustamento. Note que estamos aptos a explicar ambas as distribuições com os mesmos valores de parâmetros básicos. No presente caso, todos os físicos têm índice de citação superior a x_c e portanto, a distribuição Exponencial Estendida pode ser considerada para esta distribuição como é feito por Laherrere e Sornette (1998).

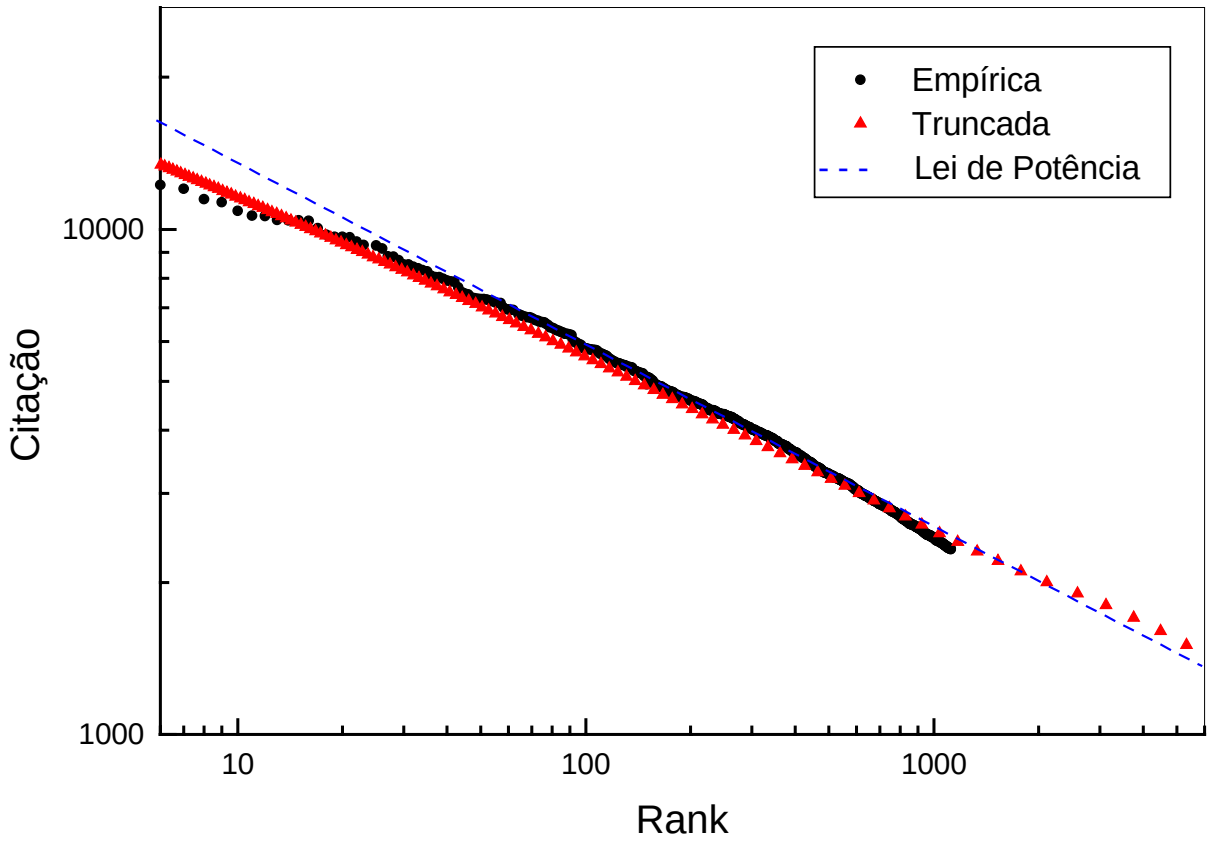

Figura 11. Gráfico log x log da dependência de Y_n como uma função do rank n, onde Y_n é o número total de citações do n-ésimo físico mais citado. Foram usados na curva teórica (gradualmente truncada) os mesmos parâmetros da Fig. 10.

Na Figura 12, fizemos um gráfico de log do rank (n) versus $(Y_n)^{\beta}$ para o caso dos físicos brasileiros. Escolhemos $\beta \cong 0.3$ que nos dá a melhor linha reta aproximada. Isto mostra que a distribuição Exponencial Estendida não é apropriada para o índice de citação de físicos brasileiros.

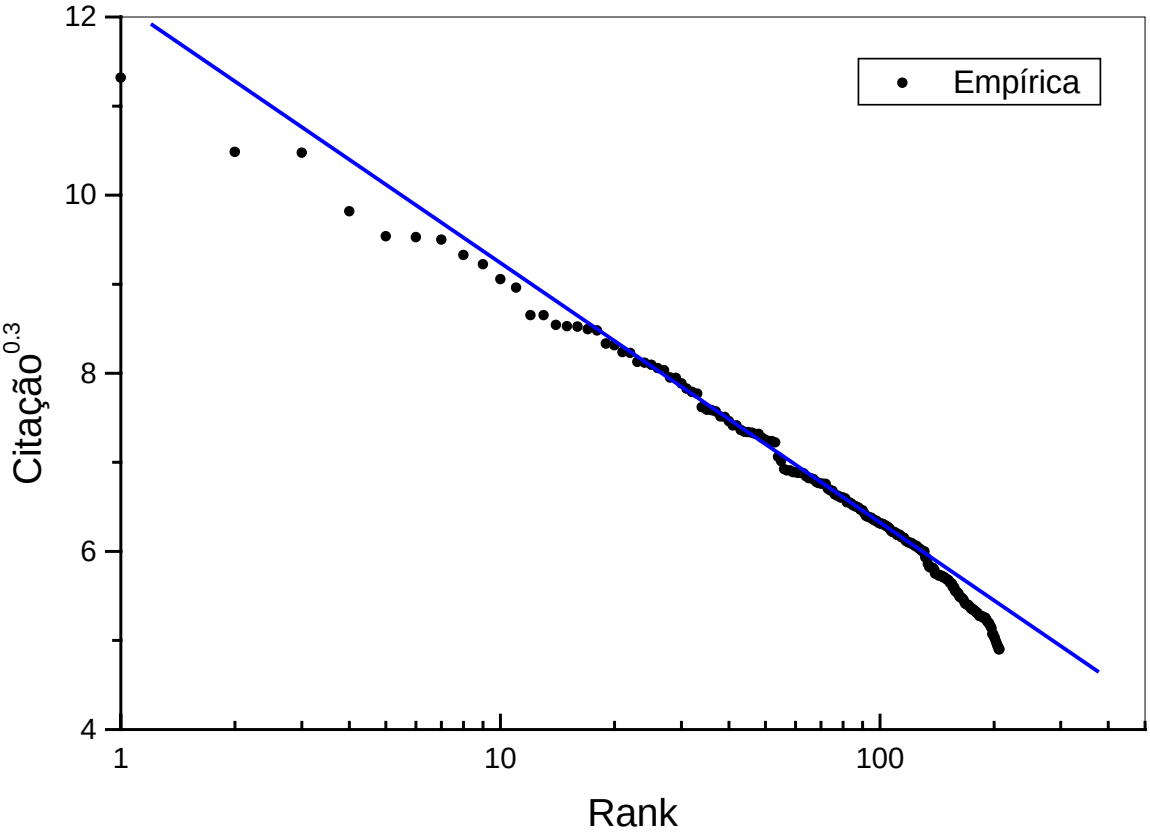

Figura 12. Log do rank (n) versus $(Y_n)^{0.3}$ para físicos brasileiros.

Observa-se que a distribuição do índice de citação de físicos é melhor dada pela Lei de Potência Gradualmente Truncada.

Na Figura 13, graficamos o número de citações (Y_n) versus o rank (n) para os 119 primeiros químicos brasileiros no ano de 1999 (Folha de São Paulo, 1999, site: http://www.uol.com.br/fsp/especial/ranking). Comparamos estes dados com a curva de Lei de Potência Gradualmente Truncada considerando α = 1.4, x_c = 3000, k = 1000, β = 0.6 e c = 2.5×10^5. O ajustamento com a lei de potência gradualmente truncada é bom.

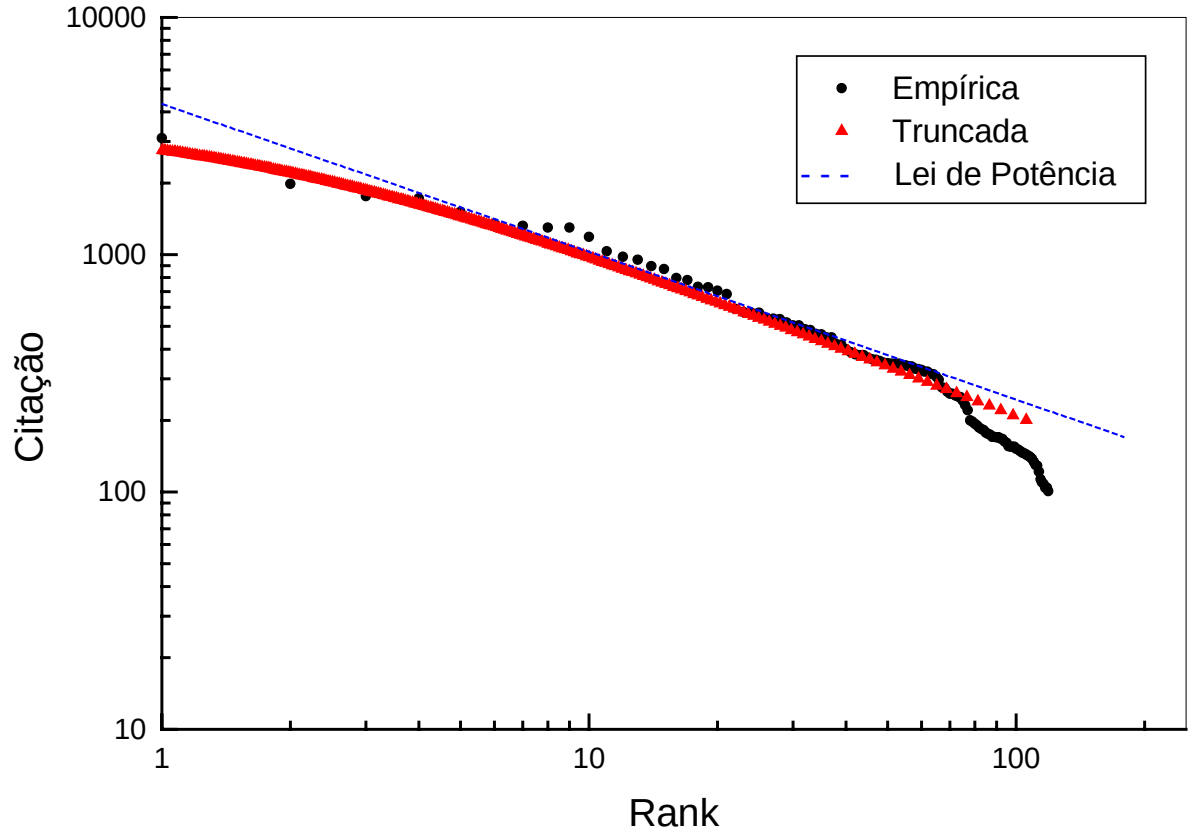

Figura 13. Gráfico de Zipf do número de citações do n-ésimo químico brasileiro do rank (Y_n) versus o rank (n) em uma dupla escala logarítmica.

Na Figura 14, fizemos o gráfico do número de citações versus o rank para os primeiros 10858 químicos (Dados do site http://pcb4122.univ-lemans.fr/chimie/chimistes.html) e comparamos este com o da distribuição de Lei de Potência Gradualmente Truncada considerando α = 1.4, x_c = 6000, k = 2000, β = 0.6 e c = 10^8. Novamente o ajustamento é bom para toda a curva.

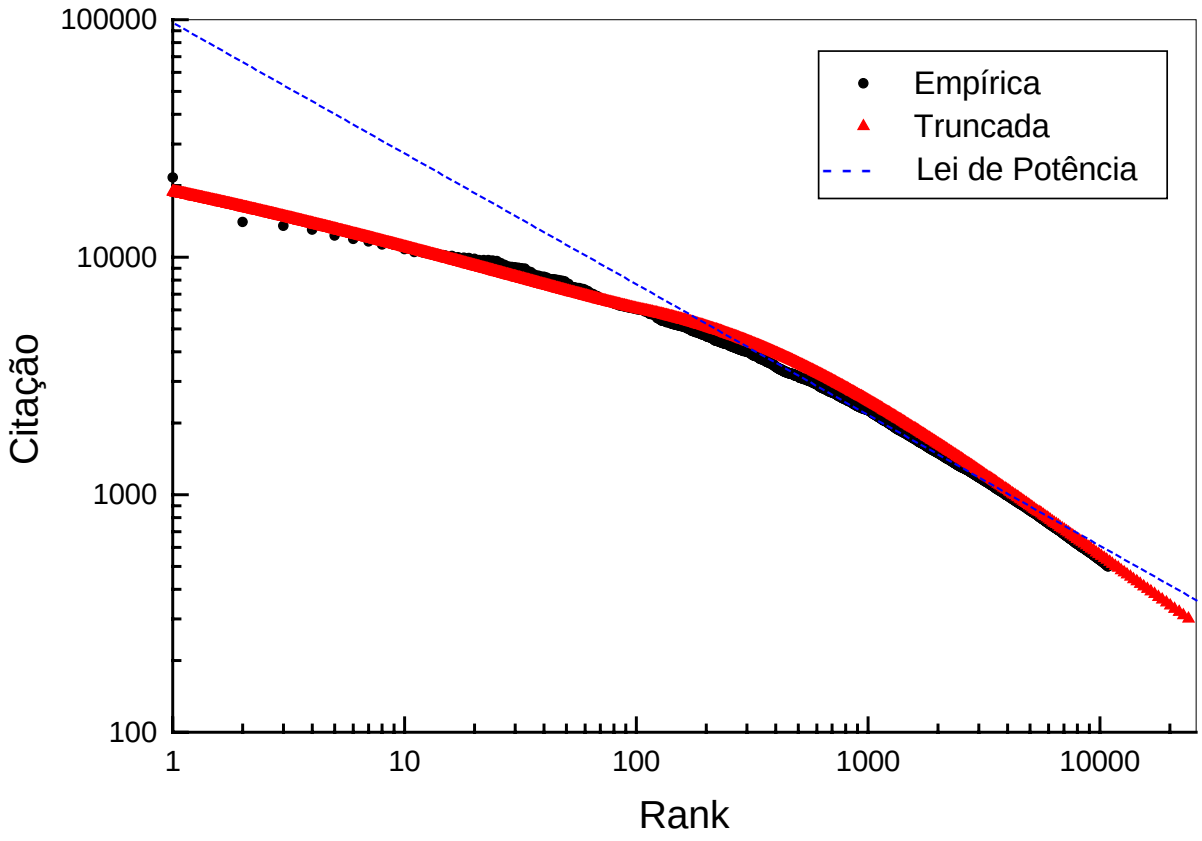

Figura 14. Gráfico de Zipf para o número de citações do n-ésimo químico do ranking internacional (Y_n) versus o rank (n) em uma dupla escala logarítmica.

Na Figura 15, mostramos um gráfico de log do rank (n) versus $(Y_n)^{\beta}$ para o caso dos químicos brasileiros. Escolhemos $\beta \cong 0.3$ que nos dá a melhor linha reta aproximada. Isto mostra que a distribuição Exponencial Estendida não é apropriada para o índice de citação dos químicos brasileiros.

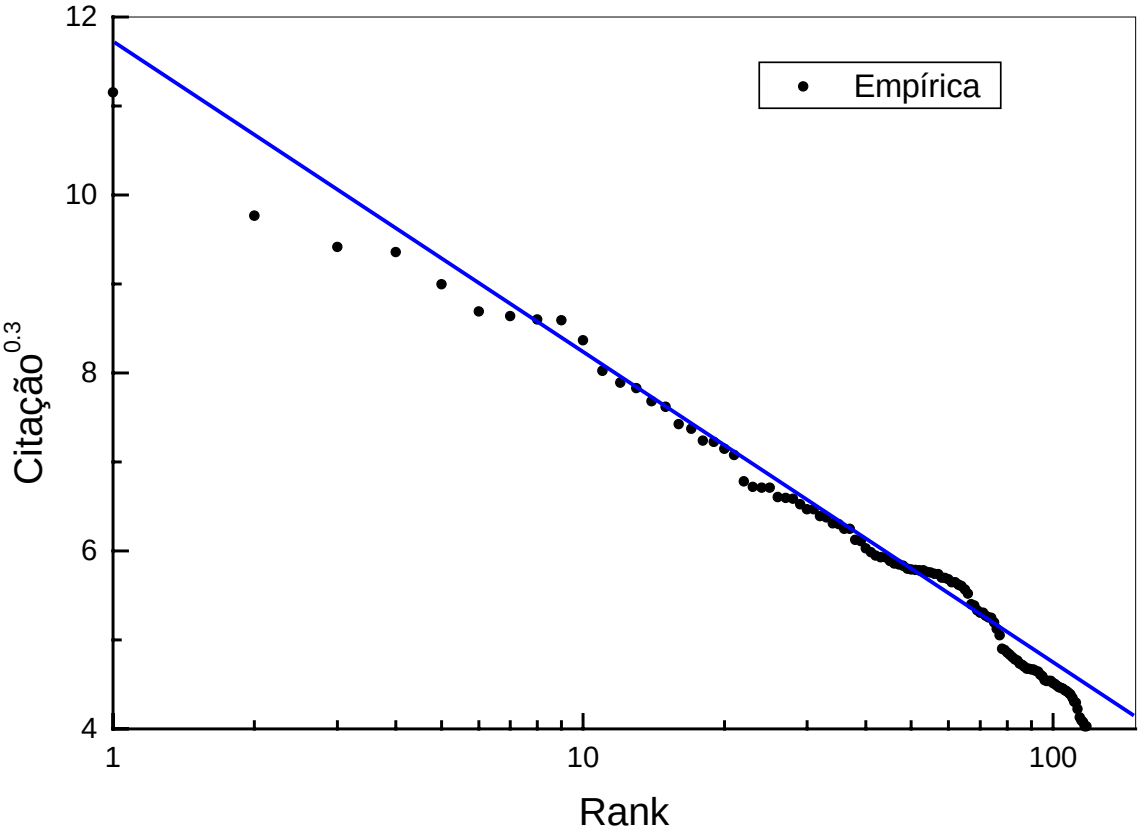

Figura 15. Log do rank (n) versus $(Y_n)^{0.3}$ para químicos brasileiros.

Na Figura 16, fizemos o gráfico de log do rank (n) versus $(Y_n)^{\beta}$ para o caso dos químicos internacionais. Da mesma forma, escolhemos $\beta \cong 0.3$ que nos dá a melhor linha reta aproximada. Isto mostra que a distribuição Exponencial Estendida, também não é apropriada para o índice de citação dos químicos internacionais.

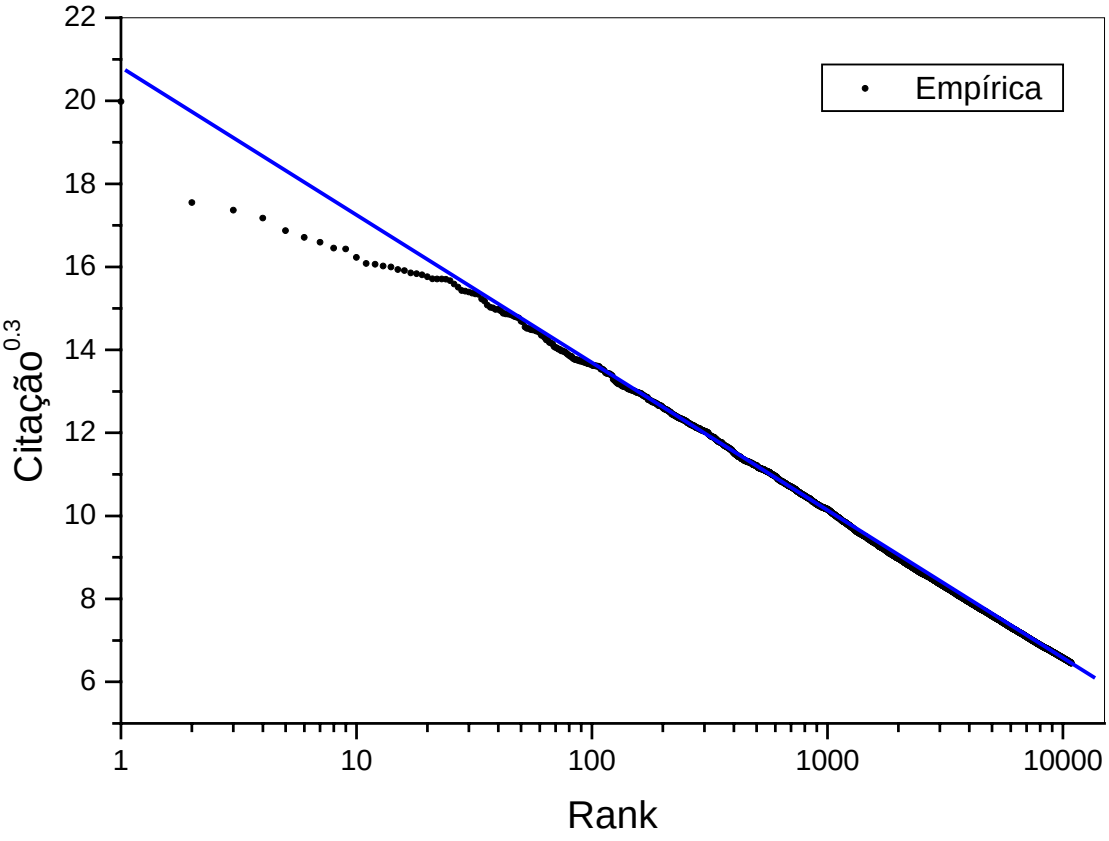

Figura 16. Log do rank (n) versus $(Y_n)^{0.3}$ para químicos internacionais.

Capítulo 6

Conclusões

6.1. Publicações Científicas

Em geral, as publicações científicas, de acordo com a sua importância, são primeiramente citadas por pessoas que trabalham no mesmo grupo, pois as mesmas estão familiarizadas com o trabalho. Mais tarde, estas publicações começam a chamar a atenção de outros grupos que trabalham na mesma área ou afins. Assim, um artigo importante que é inicialmente citado mais vezes chama a

atenção de mais pesquisadores e, portanto, faz com que a citação cresça rapidamente.

Por outro lado, artigos que são menos importantes, são inicialmente citados poucas vezes e então, chamam a atenção de apenas alguns pesquisadores tornando este menos importante, o que faz com que seu índice de citação decresça muito rapidamente. Sendo assim, a maioria dos artigos menos importantes são esquecidos rapidamente, e, somente alguns poucos artigos são citados por um longo tempo.

Isto nos dá o efeito de longa memória em termos da Estatística de Tsallis, ou realimentação positiva em termos da distribuição de Pareto. Um artigo mais citado nos estágios iniciais chama a atenção de mais pesquisadores e então é citado mais vezes, o que chama a atenção de mais pessoas ainda e é citado outras tantas vezes, e assim continua. Então o índice de citação aumenta muito mais rapidamente para artigos importantes o que nos dá uma distribuição de Lei de Potência para o índice de citação.

Através do gráfico de Zipf, observamos que os primeiros três artigos são citados mais vezes do que o esperado, através da presente distribuição. O índice de citação somente dos mais citados artigos, é esperado um crescimento com o tempo porque outros artigos são mais ou menos esquecidos em vinte anos. A presença de pontos muito fora da distribuição (outliers) no ranking dos mais citados é também chamado de efeito King e podem existir devido ao processo de amplificação (Davies, 2002; Sornette, 2002).

Nossas conclusões finais são basicamente as mesmas obtidas por Tsallis e Albuquerque (2000). Por enquanto, nós mostramos que considerando a densidade de probabilidade, obtivemos uma boa aproximação para oito graus de magnitude (10^{-4} a 10^{4}) ao invés de somente três graus de magnitude (10^{1} a 10^{4}). Isso mostra uma robustez e validade da Estatística de Tsallis neste caso.

O fato de um cientista ser citado mais vezes facilita a ele conseguir mais ajuda financeira para os seus projetos de pesquisa e melhores estudantes, o que em

troca contribui para a formação de grupos melhores e maiores. Como um artigo é inicialmente citado em artigos do mesmo grupo, qualquer artigo vindo de fora deste grupo tem mais citações num estágio inicial. Um artigo mais citado chama a atenção de outros cientistas da mesma área e é citado por outros. Um cientista mais citado pode mais eficientemente elevar seu índice de citação do que a média dos cientistas, criando mais citação e conseguindo um maior índice de citação. Então, o mecanismo de realimentação positiva aumenta a produção de qualquer parâmetro que esteja sendo analisado.

Esta realimentação positiva decresce gradualmente após certo tamanho de passo devido às limitações físicas do sistema. No caso de uma publicação científica, o índice de citação deve decrescer com o tempo depois do campo ter sido saturado ou quase todo investigado. Isto normalmente leva de 20 a 100 anos dependendo do campo. Como temos o índice de citação para somente um período de 16 anos, não observamos a Lei de Potência Gradualmente Truncada. Isto provavelmente seria observado se tivéssemos um banco de dados de um período de 30 anos ou mais.

6.2. Cientistas

No caso dos cientistas, além da limitação vinda da saturação do campo de pesquisa, a outra limitação vem da capacidade humana para trabalhar. O índice de citação de um cientista é o produto do número de seus artigos e a média da citação por artigo. Não é possível para qualquer um competir ambos em qualidade e quantidade. Alguém pode ter um grande número de artigos, mas, uma pequena média de citação por artigo, ao passo que outro cientista pode ter uma grande média de citações de um artigo, mas um pequeno número de artigos.

Isto pode ser observado claramente nas Tabelas I e II, em que listamos o número de artigos e a média de citações por artigo dos vinte físicos e químicos internacionalmente mais citados. É interessante notar que somente dois (Anderson, P. W. e Muller, K. A., no 13º. e 17º. lugar respectivamente) dentre os 20 mais citados físicos e seis (Pople, J. A., Ernst, R. R., Lehn, J. M., Smalley, R. E., Corey, E. J. e Tanaka, K. no 2º., 4º., 10º., 12º., 16º. e 20º. Lugar) dentre os 20 químicos mais

citados são premiados pelo Nobel em que a ênfase é em qualidade de um único trabalho. Embora o prêmio Nobel tenha como resultado um grande impacto, muitos deles não são cientistas mais citados por causa das poucas publicações. Então, embora não observemos a Lei de Potência Gradualmente Truncada no caso da publicação científica, isto é observado no caso dos cientistas.

Observamos que só a Lei de Potência não se ajusta para a distribuição de físicos e químicos tanto internacionais como brasileiros. A distribuição Exponencial Estendida ajusta bem somente com os 1120 físicos internacionais mais citados (Laherrere e Sornette, 1998), mas falha para os 10858 químicos internacionais mais citados. Isso acontece porque neste caso temos um grande número de químicos (10858 químicos versus somente 1120 físicos), e muitos químicos têm citações abaixo do valor crítico (x_c = 6000). Neste caso, as limitações humanas não ficam tão evidentes para todos. No caso dos físicos e químicos brasileiros, novamente a distribuição Exponencial Estendida não é válida. Somente a distribuição de Lei de Potência Gradualmente Truncada é válida para todos os casos.

Para os físicos brasileiros e internacionais, temos que os valores de todos os parâmetros α, x_c e k são os mesmos. Mas no caso dos químicos, x_c é baixo para os químicos brasileiros em comparação aos químicos internacionais. Talvez isso aconteça por causa do procedimento mais longo e burocrático para ajuda financeira no Brasil, que é um fator importante na maioria das sub-áreas da química em comparação a física, particularmente, a física teórica. É interessante notar que entre os 10 físicos mais citados no Brasil, 8 são da Física Teórica, pois estes precisam de poucos recursos financeiros para fazer o seu trabalho. Nas Tabelas III e IV, nós mostramos respectivamente, os 10 físicos e químicos mais citados no Brasil.

Observamos que as citações dos físicos brasileiros é somente 0,2% de todos os físicos. No caso dos químicos, esta porcentagem aumenta para 0,25% ($c = 2{,}5 \times 10^5$ para químicos brasileiros em relação a $c = 10^8$ para químicos internacionais). Isso mostra que o impacto de físicos e químicos brasileiros é bem menor em comparação à população que é cerca de 3% da população mundial. É preciso que se tenha mais incentivo por parte da agência governamental e privada,

para um melhor desempenho científico no Brasil, fator muito importante para o desenvolvimento de qualquer país.

Como o índice de citação de somente cientistas mais citados é válido, este não é possível distinguir entre a distribuição proposta por Tsallis e a distribuição de Lei de Potência, pois ambos nos dão o mesmo resultado neste intervalo.

Concluindo, o índice de distribuição de citação de publicações científicas e de cientistas é dado através da distribuição de Lei de Potência Gradualmente Truncada devido à longa memória e o efeito da realimentação positiva, sendo também observada em muitos sistemas econômicos e sociais. A limitação física no caso dos cientistas é dada pela limitação humana de produzir trabalhos de altíssima qualidade.

Tabela I – Citação dos 20 Físicos mais citados no período de Janeiro de 1981 a Junho de 1997.

Rank	Nome do Autor	Citação/Artigo	Artigos	Citações
1.	WITTEN E	168.37	138	23235
2.	GOSSARD AC	40.56	419	16994
3.	CAVA RJ	64.60	223	14405
4.	BATLOGG B	83.32	170	14164
5.	PLOOG K	18.95	712	13491
6.	ELLIS J	40.18	305	12255
7.	FISK Z	23.13	520	12030
8.	CARDONA M	20.08	571	11465
9.	NANOPOULOS DV	38.61	293	11314
10.	HEEGER AJ	33.98	320	10872
11.	LEE PA	72.89	146	10642
12.	SUZUKI T	7.58	1401	10617
13.*	ANDERSON PW (1977)	80.30	138	10439
14.	SUZUKI M	11.60	898	10417
15.	FREEMAN AJ	26.76	389	10411
16.	TANAKA S	10.80	963	10404
17.*	MULLER KA (1987)	82.37	122	10049
18.	SCHNEEMEYER LF	62.62	156	9768
19.	CHEMLA DS	59.68	162	9668
20.	MORKOC H	20.27	477	9668

* Prêmio Nobel

Tabela II – Citação dos 20 Químicos mais citados no período de Janeiro de 1981 a Junho de 1997.

Rank	Nome do Autor	Citação/Artigo	Artigos	Citações
1.	BAX A	142.47	152	21655
2.*	POPLE JA (1998)	79.80	176	14044
3.	SCHLEYER PV	25.83	525	13559
4.*	ERNST RR (1991)	71.81	182	13069
5.	WHITESIDES GM	38.71	318	12310
6.	SCHAEFER HF	23.15	515	11921
7.	HUFFMAN JC	20.20	577	11654
8.	RHEINGOLD AL	13.63	830	11317
9.	SEEBACH D	32.31	349	11275
10.*	LEHN JM (1987)	35.25	307	10823
11.	MEYER TJ	39.29	267	10490
12.*	SMALLEY RE (1996)	108.92	96	10456
13.	BARD AJ	31.13	333	10365
14.	TRUHLAR DG	31.43	328	10310
15.	STEWART JJP	261.00	39	10179
16.*	COREY EJ (1990)	33.43	303	10129
17.	YAMAMOTO Y	10.70	935	10007
18.	TANAKA T	10.44	954	9961
19.	COTTON FA	15.63	634	9911
20.*	TANAKA K (1959)	8.17	1202	9820

* Prêmio Nobel

Tabela III – Citação dos 10 Físicos Brasileiros mais citados (citações até Agosto de 1999).

Rank	Nome do Autor	Instituição	Citações
1.	Amir Ordacgi Caldeira	IFGW/Unicamp	3.258
2.	Constantino Tsallis	CBPF	2.525
3.	Rogério C. de Cerqueira Leite	IFGW/Unicamp	2.516
4.	José Antonio Brum	IFGW/Unicamp	2.027
5.	Raul José Donangelo	IF/UFRJ	1.840
6.	Carlos Henrique de Brito Cruz	IFGW/Unicamp	1.833
7.	Eduardo Luiz Damiani Bica	IF/UFRGS	1.817
8.	Fernando Cerdeira	IFGW/Unicamp	1.709
9.	Carlos Augusto Bertulani	IF/UFRJ	1.549
10.	Luiz Carlos Moura Miranda	CCE/UEM	1.496

Tabela IV – Citação dos 10 Químicos Brasileiros mais citados (citações até Agosto de 1999).

Rank	Nome do Autor	Instituição	Citações
1.	Otto Richard Gottlieb Fiocruz	Fund. Oswaldo Cruz	3.099
2.	Elias A. Guidetti Zagatto Cena	CENA/USP/Piracicaba	1.991
3.	Henrique Elise Toma	IQ/USP/São Paulo	1.763
4.	Frank Herbert Quina	IQ/USP/São Paulo	1.725
5.	Faruk Jose Nome Aguilera	CCFM/UFSC	1.513
6.	Francisco Jose Krug	CENA/USP/Piracicaba	1.350
7.	Boaventura Freirre dos Reis	CENA/USP/Piracicaba	1.322
8.	Nicola Petragnani	IQ/USP/São Paulo	1.302
9.	Teresa B. Iwasita de Vielstich	IQ/USP/São Carlos	1.299
10.	João Valdir Comasseto	IQ/USP/São Paulo	1.190

7. Referências Bibliográficas

Addison, Paul S. Fractals and Chaos – An illustrated course. IOP Publishing Ltd., 1997.

Ahrens, L. H. The log-normal distribution of the elements (A fundamental law of geochemistry and its subsidiary). Geochimica et Cosmochimica Acta **5**, p. 49-73, 1954.

Bagrow, J. P., Rozenfeld, H. D., Bollt, E. M. and Ben-Avraham, D. How famous is a scientist? – Famous to those who know us. Europhys. Lett. **67** (4), p. 511-516, 2004.

Bak, P. How Nature Works. Oxford University Press, Oxford, 1997.

Bassingthwaighte, J. B., Liebovitch, L. S. and West, B. J. Fractal Physiology. Oxford Univ. Press, New York, 1994.

Baur, P. Log-normal distribution of water permeability and organic solute mobility in plant cuticles. Plant, Cell and Environment **20**, p. 167-177, 1997.

Biondini, R. Cloud motion and rainfall statistics. Journal of Applied Meteorology **15**, p. 205-224, 1976.

Boag, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. Journal of the Royal Statistical Society B **11**, p. 15-53, 1949.

Capra, F. A teia da vida. Uma Nova Compreensão Científica dos Sistemas Vivos. Editora Cultrix, São Paulo, 1996.

Chabaud B., Naert, A., Peinke, J., Chillà, F., Castaing, B. and Hébral, B. Transition Toward Developed Turbulence. Phys. Rev. Lett. **73**, p. 3227-3230, 1994.

Davies, J. A. The individual success of musicians, like that of physicists, follows a stretched exponential distribution. Eur. Phys. J. B **27**, p. 445-447, 2002.

De Moivre, A. The Doctrine of Chances: or, a Method of Calculating the Probability of Events in Play. W. Pearson, 3nd. edition, Woodfall, London, 1733.

Di Giorgio, C., Krempff, A., Guiraud. H., Binder, P., Tiret, C. and Dumenil, G. Atmospheric pollution by airborne microorganisms in the City of Marseilles. Atmospheric Environment **30**, p. 155-160, 1996.

Einstein, A. and Infield, L. (1966). The Evolution of Physics From Early Concepts to Relativity and Quanta. Simeon and Schuster, New York,1966.

Galambos, J. The Assymptotic Teory of Extreme Order Statistics. John Wiley & Sons, New York, 1978.

Gauss, K. F. Theoria motus corporum coelestium in sectionibus conicis Solem ambientium. F. Perthes and I.H. Besser, Hamburg, Germany, 1809.

Geller, R. J. Earthquake prediction: a critical review. Geophysical Journal International **131**, p. 425-450, 1997.

Gleria, I., Matsushita, R. e Silva, S. da. Sistemas Complexos, criticalidade e leis de potência. Revista Brasileira de Ensino de Física, v. 26, n. 2, p. 99-108, 2004.

Gupta, H. M., Campanha, J. R. and Chavarette, F. R. Power Law Distribution in Education: Effect of Economical Teaching and Study Conditions in University Entrance Examination. Int. Journal of Modern Physics C, **14**, p. 449-457, 2003.

Gupta, H. M., Campanha, J. R. and Prado, F. D. Power Law Distribution in Education: University Entrance Examination. Int. Journal of Modern Physics C, **11**, p. 1273-1279, 2000.

Gupta, H. M., Campanha, J. R. The gradually truncated Lévy flight: stochastic process for complex systems. Physica A **275**, p. 531-543, 2000.

Herdan, G. The relation between the dictionary distribution and the occurrence distribution of word length and its importance for the study of quantitative linguistics. Biometrika **45**, p. 222-228, 1958.

Hurst, H. E. Long-term storage capacity of reservoirs. Transactions of the American Society of Civil Enginners, **116**, pp. 770-799, 1951.

Kondo, K. The log-normal distribution of the incubation time of exogenous diseases. Japanese Journal of Human Genetics **21**, p. 217-237, 1977.

Laherrere, J., Sornette, D. Stretched exponential distributions in nature and economy: “fat tails” with characteristic scales. Eur. Phys. J. B **2**, p. 525-539, 1998.

Laplace, P. S. Mémoire sur les probabilités. Mémoirs de l’Académie Royale des Sciences de Paris, 1781.

Lévy, J. S. War in the Modern Great Power System 1495-1975. Lexington: University of Kentucky Press, 1983.

Lévy, P. Théorie de l’Addition des Variables Aléatories. Gauthier-Villars, Paris, 1937.

Limpert, E., Abbt, M., Asper, R., Graber, W. K. Godet, F., Stahel, W. A. and Windhab, E. J. Life is log normal. Keys and clues to understand patterns of multiplicative interactions from the disciplinary to the transdisciplinary level. Technology and Society, Zurick, p. 20-24, 2000.

Macau, E. E. N. Sistemas Complexos. Anais do I Congresso de Dinâmica e Aplicações, Rio Claro, v. 1, p. 29-49, 2002.

Magurran, A. E. Ecological Diversity and its Measurement. Croom Helm, London, 1988.

Malamud, B., Morein, G. and Turcotte, D. Forest fires: an exemple of self-organized critical behaviour. Science **281**, p. 1840-2, 1998.

Malanca, A. Gaidolfi, L., Pessina, V. and Dallara, G. Distribution of 226-Ra, 232-Th, and 40-K in soils of Rio Grande do Norte (Brazil). Journal of Environmental Radioactivity **30**, p. 55-67, 1996.

Mandelbrot, B. The Variation of Certain Speculative Prices. Journal of Business **36**, p. 394-419, 1963.

Meadows, A. J. A comunicação Científica. Brasília: Briquet de Lemos, 2000.

Ott, A., Bouchard, J. P., Langevin, D. and Urbach, W. Anomalous Diffusion in “Living Polymers”: A Genuine Lévy Flight? Phys. Rev. Lett. **65**, p. 2201-2204, 1990.

Pareto, V. Cours d’Économie Politique. Reprinted as a volume of Oeuvres Complétes, Droz, Geneva, 1896-1965.

Peng, C. K., Mietus, J., Hausdorff, J. M., Havlin, S., Stanley, H. E. and Goldberger, A. L. Long Range Anticorrelations and Non-Gaussian Behavior of the Heartbeat. Phys. Rev. Lett. **70**, p. 1343-1346, 1993.

Preston, F. W. Pseudo-log-normal distributions. Ecology **62**, p. 355-364, 1981.

Razumovsky, N. K. Distribution of metal values in ore deposits. Comptes Rendus (Doklady) de l’Académie des Sciences de l’URSS **9**, p. 814-816, 1940.

Redner, S. How popular is your paper? An empirical study of the citation distribution. Eur. Phys. J. B **4**,131, 1998.

Redner, S. Citation Statistics From 110 Years of Physcal Review. arXiv:physics/0506056, v. 2, jun/2005.

Reinders, R. D., Evers, E. G., Jonge, R. de e van Leusden F. M. Variation in the numbers of Shiga toxin – producing *Escherichia coli* O157 in minced beef. RIVM, 2002.

Romero, R. A. and Sutton, T. B. Sensitivity of Mycosphaerella fijiensis, causal agent of black sigatoka of banana, to propiconozole. Phytopathology **87**, p. 96-100, 1997.

Santos, R. N. M. Produção Científica: Por que medir? O que medir? Revista Digital de Biblioteconomia e Ciência da Informação, Campinas, v. 1, pg. 22-38, jul/dez de 2003.

Sartwell, P. E. The distribution of incubation periods of infectious disease. American Journal of Hygiene **51**, p. 310-318, 1950.

Sartwell, P. E. The incubation period and the dynamics of infectious disease. American Journal of Epidemiology **83**, jp. 204-216, 1966.

Sartwell, P. E. The incubation period of poliomyelitis. American Journal of Public Health and the Nation's Health **42**, p. 1403-1408, 1952.

Shockley, W. On the statistics of individual variations of productivity in research laboratories. Proc IRE **45**, p. 279-290, 1957.

Simkin, M. V. and Roychowdhury, V. P. Read before you cite. Complex Systems **14**, 269, 2003.

Simkin, M. V. and Roychowdhury, V. P. Stochastic modeling of citation slips. Scientometrics **62**, p. 367-384, 2005.

Solomon, T. H. Weeks, E. R. and Swinney, H. L. Observation of Anomalous Diffusion and Lévy Flights in a Two-Dimensional Rotating Flow. Phys. Rev. Lett. **71**, p. 3975-3978, 1993.

Sornette, D. Predictability of catastrophic events: Material rupture, earthquakes, turbulence, financial crashes, and human birth. Proc. Natl. Acad. Sci. USA **99**, p. 2522-2529, 2002.

Sugihara, G. Minimal community structure: An explanation of species abundance patterns. American Naturalist **116**, p. 770-786, 1980.

Tsallis, C. and Albuquerque M. P. Are citations of scientific papers a case of nonextensivity? Eur. Phys. J. B **13**, p. 777-780, 2000.

Tsallis, C. Possible Generalization of Boltzmann-Gibbs Statistics. J. Stat. Phys. **52**, p. 479-487, 1998.

Tsallis, C. Nonextensive Statistics: Theoretical, Experimental e Computational Evidences and Connections. Brazilian Journal of Physics **29**, p. 1-35, 1999.

Vines, G. Mass extinctions, in Inside Science. New Scientist **11**, December, p. 1-4, 1999.

Williams, C. B. A note on the statistical analysis of sentence length as a criterion of literary style. Biometrika **31**, p. 356-361, 1940.

Wilson, M.A., Hoff, W.D., Hall, C., McKay, B., Hiley, A. Kinetics of moisture expansion in fired clay ceramics: a (Time)1/4 law. Physical Review Letters, 90, 12, 2003.

Zimba, H. F. e Mueller, S. P. M. Colaboração Internacional e Visibilidade Científica de Países em Desenvolvimento. Informação e Sociedade: Estudos, v. 14, n. 1, 2004.

Zipf, G. K. Human Behavior and the Principle of Least Effort. Addison-Wesley, 1949.

8. Apêndice

Artigo Aceito para Publicação:

Gupta, H. M., Campanha, J. R., and Pesce, R. A. G. Power-Law Distributions for the Citation Index of Scientific Publications and Scientists. Brazilian Journal of Physics, vol. 35, n. 4, December, 2005.