インターネット計測とデータ解析 第 5 回

長 健二朗

2013 年 5 月 8 日

前回のおさらい

第 4 回 分布と信頼区間 (5/1)

- 正規分布
- 信頼区間と検定
- 分布の生成
- 演習: 信頼区間
- 課題 **1**

今日のテーマ

第 5 回 多様性と複雑さ

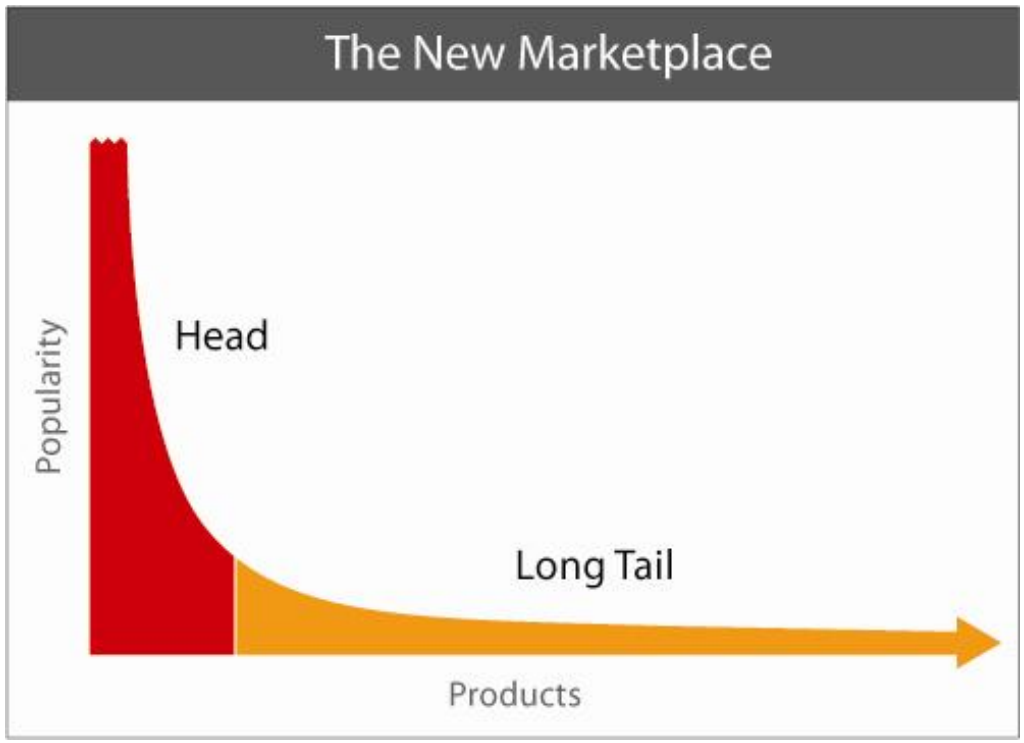
- ロングテール
- Web アクセスとコンテンツ分布
- べき乗則と複雑系
- 演習: べき乗則解析

ロングテール

オンライン小売サービスのビジネスモデル

- ヘッド: 少数の売れ筋商品、リアルの店舗の守備範囲
- テール: 多様な売上下位商品、オンライン店舗の売上の特徴

いまでは多様なニッチマーケットを指す言葉として広く使われる



source: http://longtail.com/

複雑さ

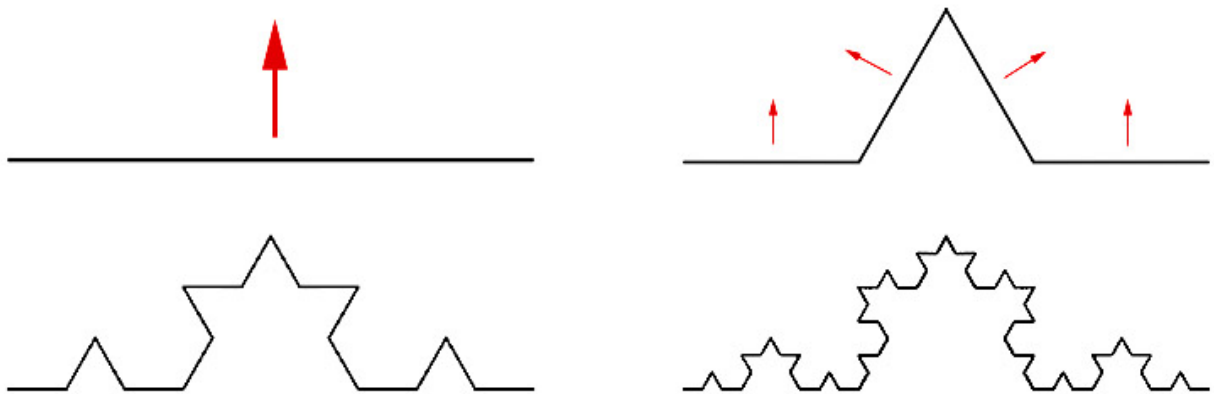
複雑さの科学

- 多数の因子が相互に影響して複雑な挙動を示すシステム
  - カオス、フラクタル、非線形力学など
- 世界は複雑系に満ちている
- 従来の還元主義的手法で解析が困難
  - 複雑な現象を複雑なまま理解する必要
- 90 年台から盛んに研究
  - 還元主義的手法で解ける未解決な問題が減ってきた
  - コンピュータによる解析やシミュレーション

べき乗則と複雑系

べき乗則

- べき乗則は複雑系を示す特徴のひとつ
  - べき乗則: 観測量がパラメータのべき乗に比例
  - 自己相似的 (フラクタル)
- さまざまな自然現象、社会現象、インターネットサービスで観測される
- スケールフリー: 特徴的なスケールを持たない



コッホ曲線の作成：海岸線に似たフラクタル図形

ジフ (Zipf) の法則

- 1930 年代に順位付けされたデータの出現頻度で発見された経験則
- シェアは順位に反比例
  - 出現頻度が k 番目に大きい要素が占める割合が $1/k$ に比例
- 社会科学や自然科学、データ通信でさまざまな現象が確認される
  - 英単語の出現頻度、都市の人口、富の分配など
  - ファイルサイズ、ネットワークトラフィックなど
- リニアスケールのグラフではロングテール、ログログスケールのグラフではヘビーテイルになる

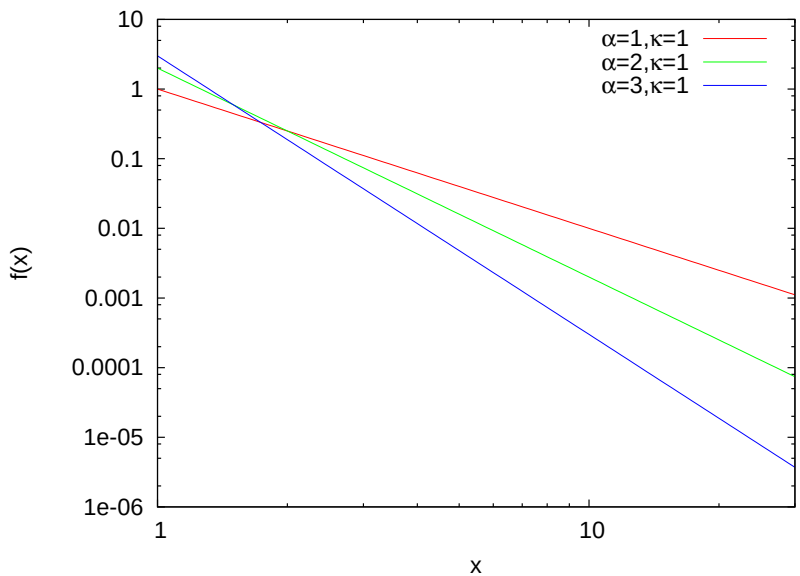
べき分布 (power-law distribution)

- べき分布: ある量が観測される確率がその大きさのべき乗に比例

$$f(x) = ax^k$$

- 両対数グラフに書くと線形になる

$$\log f(x) = k \log x + \log a$$



インターネットの複雑さ

トポロジーの複雑さ

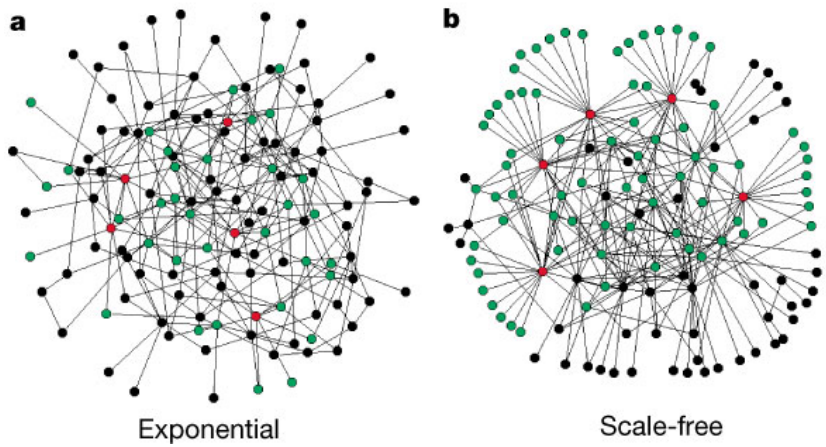
- スケールフリー: ノードの次数にべき乗則の偏り
  - 多数の小次数ノードと少数の大次数ノード
  - 平均的なサイズがない
- スモールワールド:
  - コンパクト: 任意のノード間の距離は短い
  - クラスタ: 友達の友達は友達

トラフィックの挙動 (時系列解析)

- 自己相似性
- 長期依存性

スケールフリーネットワーク

- ネットワークノードの次数分布がべき乗則に従う
  - ほとんどのノードの次数は 1 や 2
  - 一部のノードの次数は桁違いに大きい (ハブノード)
- スモールワールド
  - ハブノードを経由して任意のノードに短距離で到達
  - ハブノードは情報の拡散機能 (感染症の場合も)
- 発生の仕組み: preferential attachment: rich get richer
  - 次数の大きいノードに接続しやすい仕組み
- 耐故障性、耐攻撃性
  - ランダムなノードの故障には強い
  - ハブノードへの攻撃には弱い



source: Error and attack tolerance of complex networks. R. Albert, H. Jeong, A. Barabasi. Nature 406, July 2000.

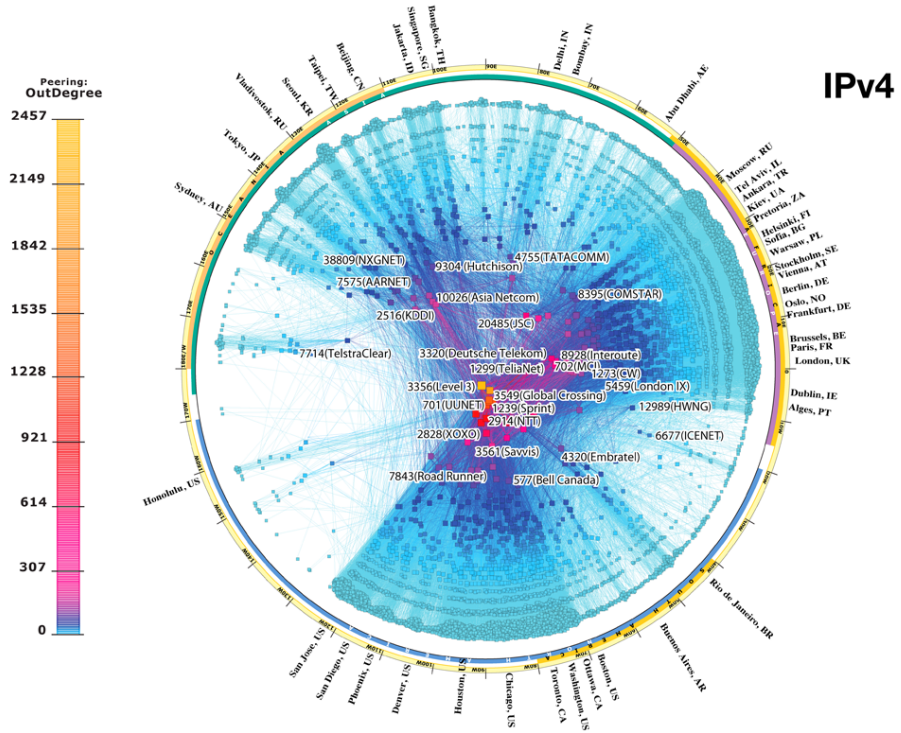
インターネットの AS 構造の例

CAIDA AS CORE MAP 2009/03

- AS の登録都市の経度、AS の out-degree

IPv4
INTERNET TOPOLOGY MAP
JANUARY 2009

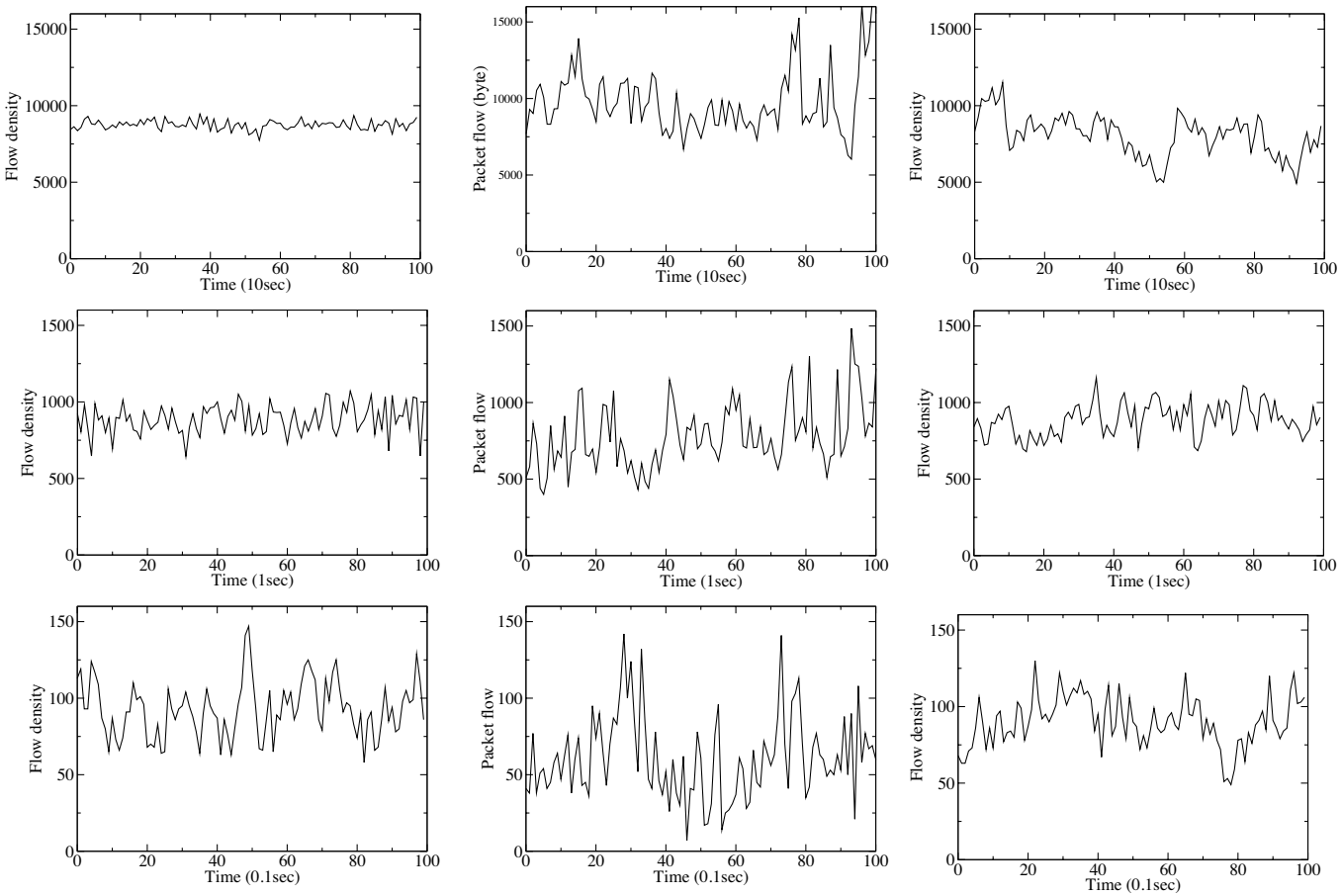
AS-level INTERNET GRAPH



http://www.caida.org/research/topology/as_core_network/

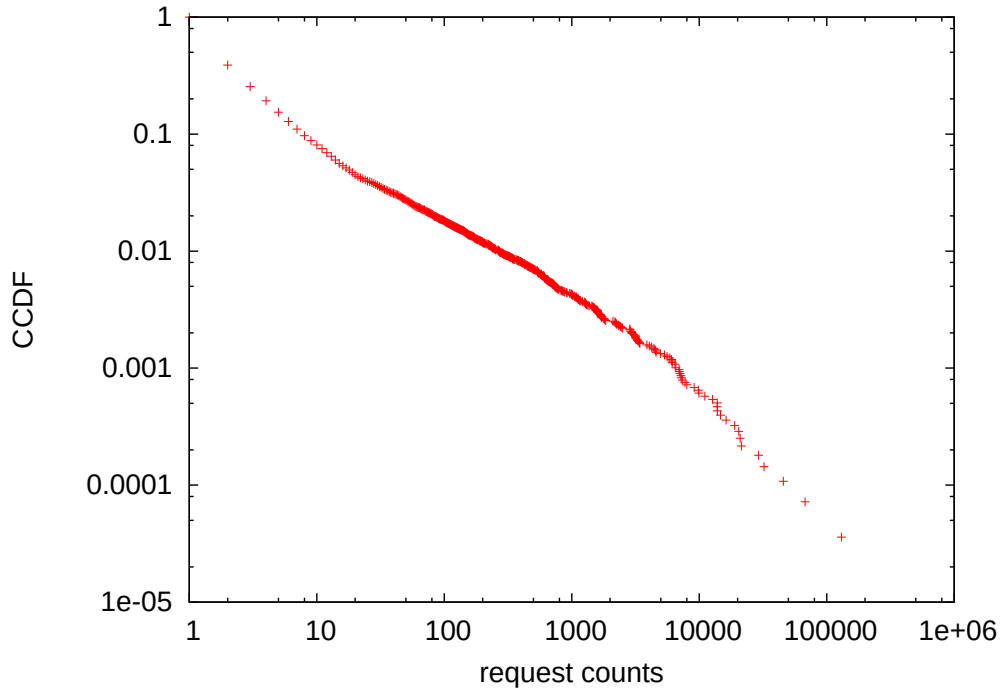
ネットワークトラフィックの自己相似性

- (左) 指数関数モデル (中) 実トラフィック (右) 自己相似モデル
- 時間粒度: (上)10sec (中)1 sec (下)0.1 sec



Web アクセスとコンテンツ分布

- Web の世界にもいたるところに、べき分布が存在
  - Web ページの被リンク数やアクセス数、検索キーワード



JAIST サーバのコンテンツ毎のアクセス数分布

さまざまな分布

- 二項分布
- ポアソン分布
- 正規分布
- 指数分布
- べき分布

二項分布 (binomial distribution)

- ベルヌーイ試行 (bernoulli trial): 試行の結果が 2 種類しかない試行
- 1 回の試行の成功率を p とし、n 回の試行の成功数 k の離散確率分布

確率関数 (PDF)

$$P[X = k] = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$$

ここで

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

$$mean : E[X] = np, variance : Var[X] = np(1-p)$$

二項分布は n が大きくなるとポアソン分布で近似できる

ポアソン分布 (poisson distribution)

まれにしか発生しない事象の時間あたりの発生回数はポアソン分布に従う

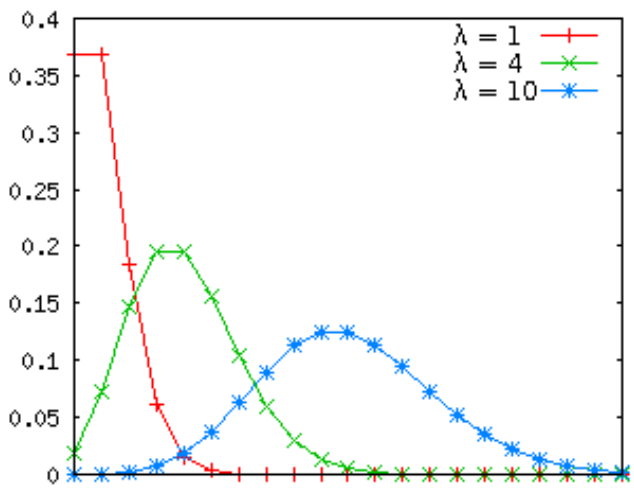
- 交通事故死亡者数や、遺伝子の突然変異数など

ポアソン分布はただひとつの平均パラメータ $\lambda > 0$ で表される

確率関数 (PDF)

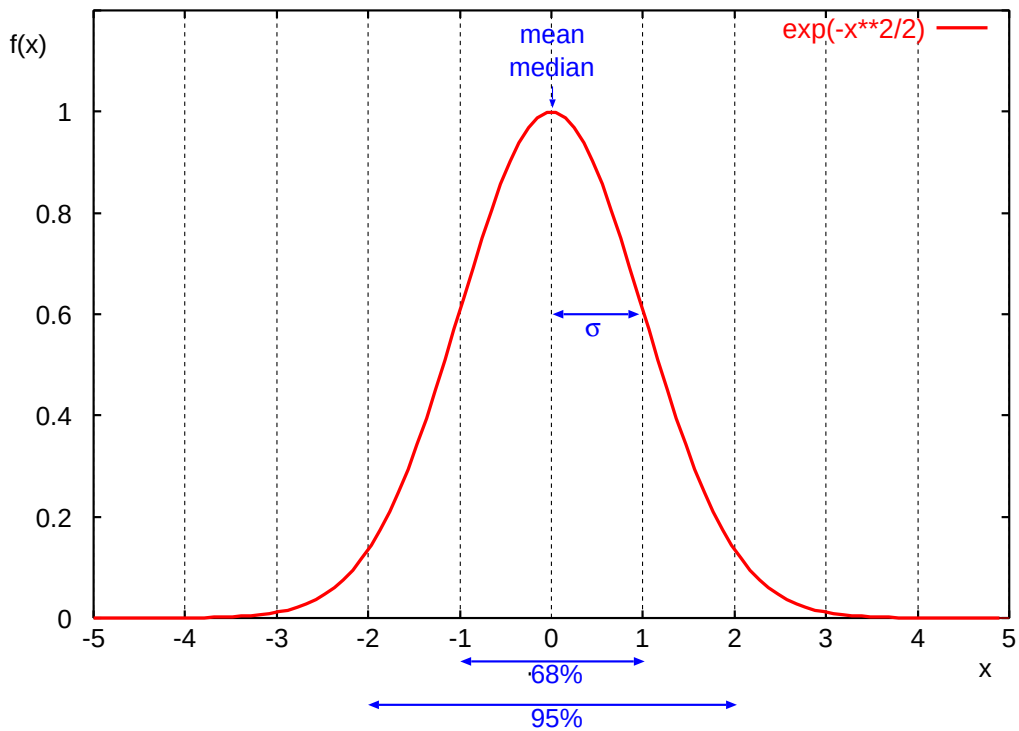
$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$mean : E[X] = \lambda, variance : Var[X] = \lambda$$



正規分布 (normal distribution) 1/2

- つりがね型の分布、ガウス分布とも呼ばれる
- 2 つの変数で定義: 平均 μ、分散 σ^2
- 乱数の和は正規分布に従う
- 標準正規分布: $\mu = 0, \sigma = 1$
- 正規分布ではデータの
  - 68%は $(mean \pm stddev)$
  - 95%は $(mean \pm 2stddev)$ の範囲に入る



正規分布 (normal distribution) 2/2

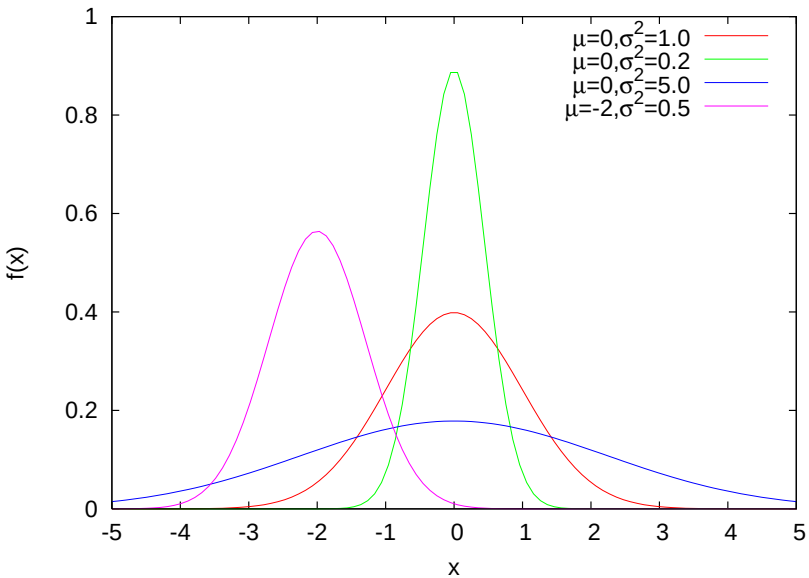
確率密度関数 (PDF)

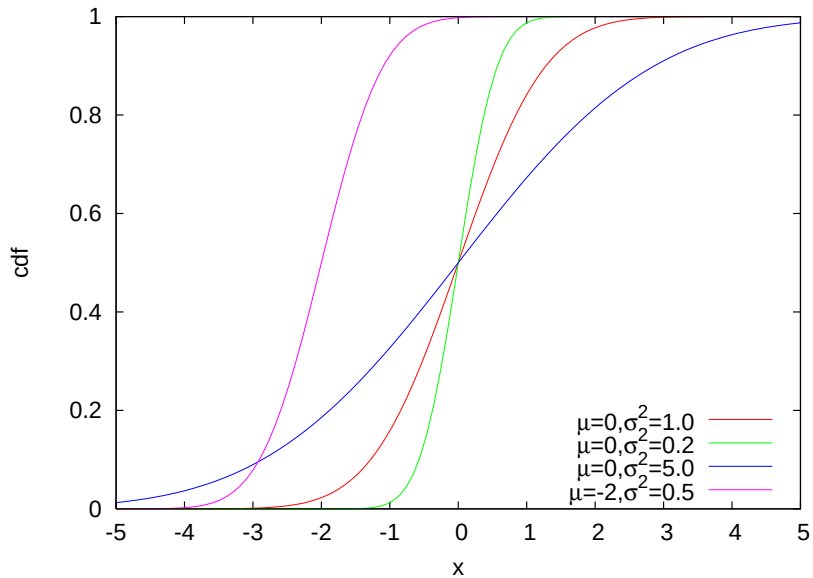
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

累積分布関数 (CDF)

$$F(x) = \frac{1}{2}(1 + erf\frac{x-\mu}{\sigma\sqrt{2}})$$

$$\mu : mean, \sigma^2 : variance$$





指数分布 (exponential distribution)

一定の確率で発生する独立事象の発生間隔は指数分布に従う

- 電話の発呼間隔や、TCP セッションの発生間隔など

確率密度関数 (PDF)

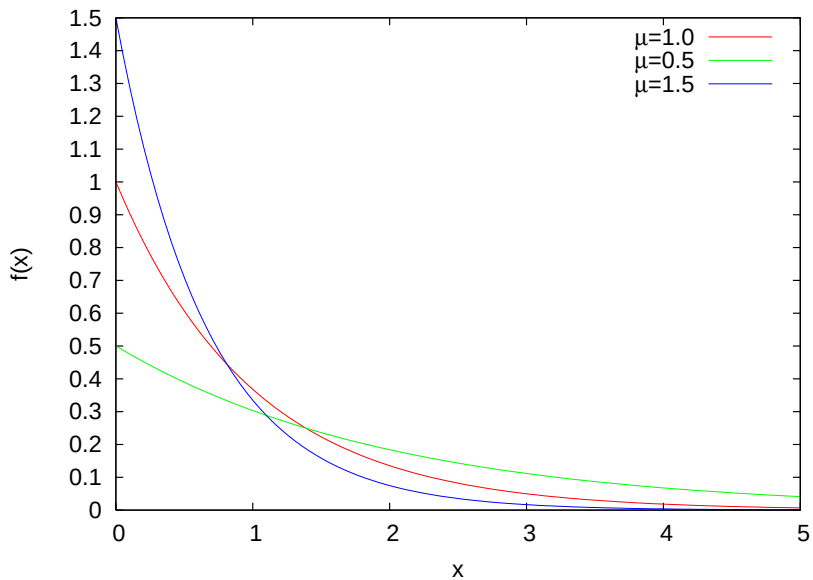
$$f(x) = \lambda e^{-\lambda x}, (x \geq 0)$$

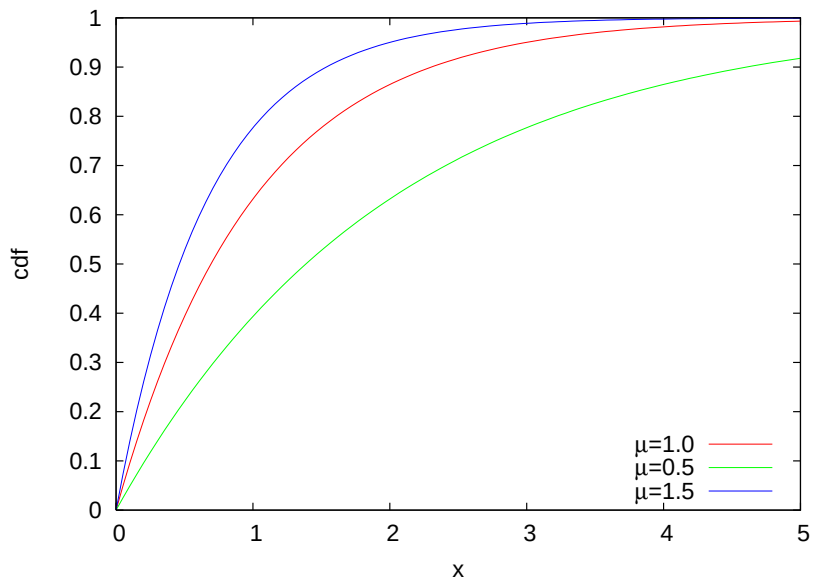
累積分布関数 (CDF)

$$F(x) = 1 - e^{-\lambda x}$$

$$\lambda > 0 : rate\ parameter$$

$$mean : E[X] = 1/\lambda, variance : Var[X] = \lambda^{-2}$$





パレート分布 (pareto distribution)

パレート分布: ネットワーク研究で最も使われる べき分布

確率密度関数 (PDF)

$$f(x) = \frac{\alpha}{\kappa}(\frac{\kappa}{x})^{\alpha+1}, (x > \kappa, \alpha > 0)$$

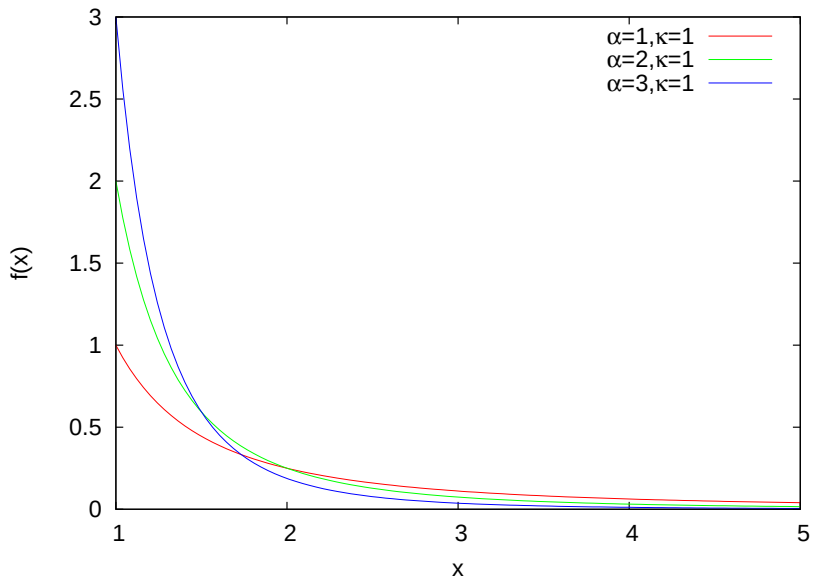
累積分布関数 (CDF)

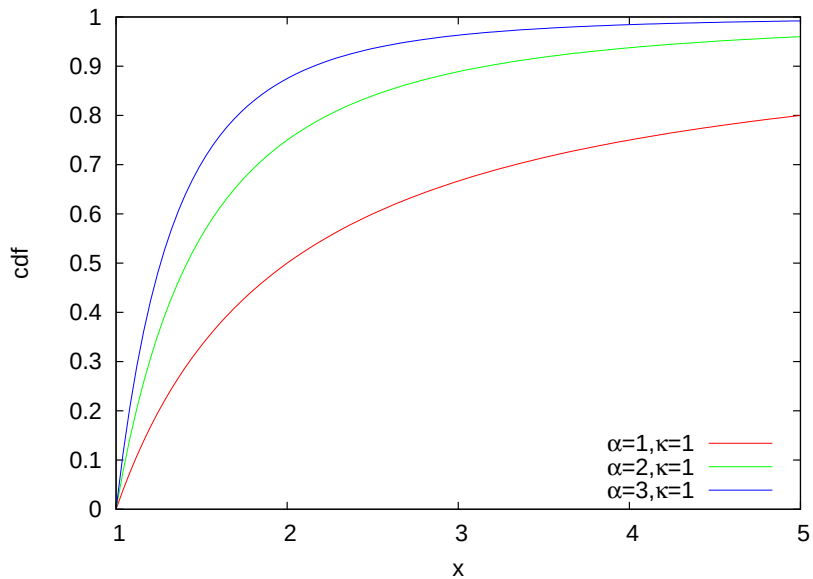
$$F(x) = 1 - (\frac{\kappa}{x})^{\alpha}$$

$$\kappa : minimum\ value\ of\ x, \alpha : pareto\ index$$

$$mean : E[X] = \frac{\alpha}{\alpha - 1}\kappa, (\alpha > 1)$$

if $\alpha \leq 2$, variance $\to \infty$. if $\alpha \leq 1$, mean and variance $\to \infty$.





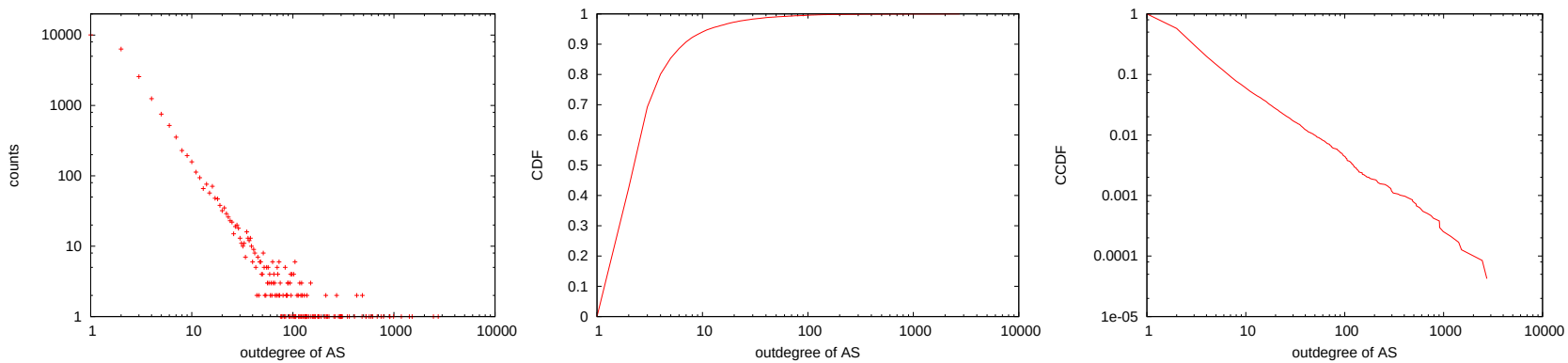
相補累積分布関数 (CCDF)

Complementary Cumulative Distribution Function (CCDF)
べき分布は分布のテイル部分 (値の大きい要素) に特徴

CCDF: x より大きい値の合計が全体に占める割合

$$F(x) = 1 - P[X <= x]$$

- CCDF はログログスケールで描画
  - テイル部分の分布や、スケールフリーな性質を見る



次数分布 (左) CDF(中) CCDF(右)

CCDF のプロット

CDF のプロット

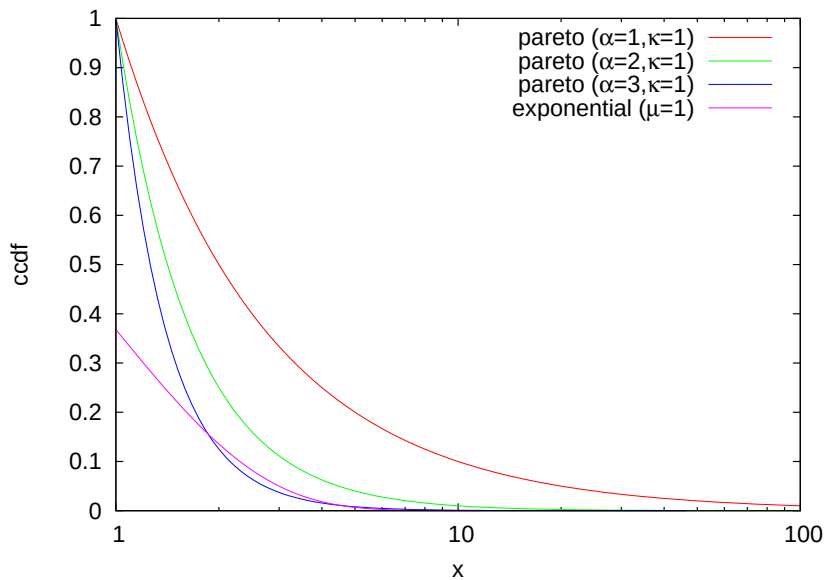
- $x_i, i \in \{1, \ldots, n\}$ を値順にソート
- $(x_i, \frac{1}{n} \sum_{k=1}^{i} k)$ をプロット
- Y 軸は通常リニアスケール

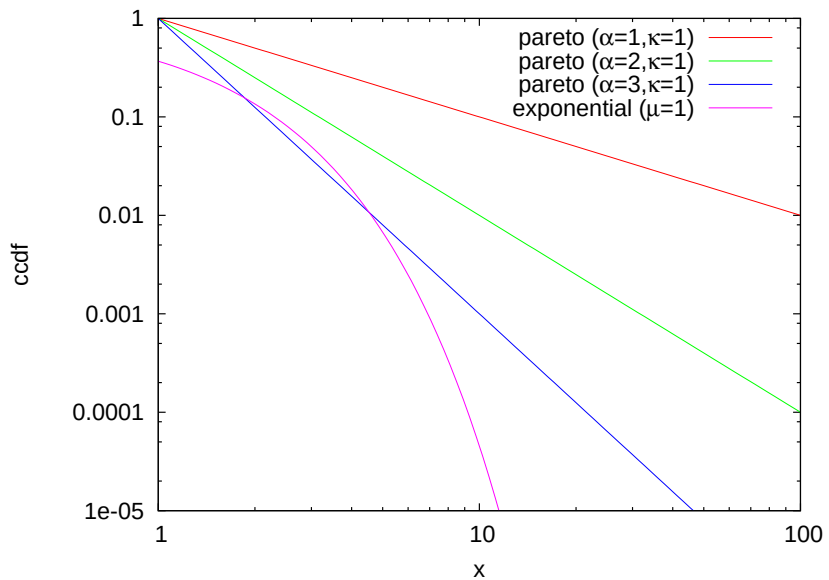
CCDF のプロット

- $x_i, i \in \{1, \ldots, n\}$ を値順にソート
- $(x_i, 1 - \frac{1}{n} \sum_{k=1}^{i-1} k)$ をプロット
- 通常 XY 軸ともログスケール

パレート分布の CCDF

- log-linear (左)
  - 指数分布が直線
- log-log (右)
  - パレート分布が直線





前回の演習: 正規乱数の生成

- 正規分布に従う疑似乱数の生成
  - 一様分布の疑似乱数生成関数 (ruby の rand など) を使って、平均 u、標準偏差 s を持つ疑似乱数生成プログラムを作成
- ヒストグラムの作成
  - 標準正規分布に従う疑似乱数を生成し、そのヒストグラム作成、標準正規分布であることを確認する
- 信頼区間の計算
  - サンプル数によって信頼区間が変化することを確認
    疑似正規乱数生成プログラムを用いて、平均 60, 標準偏差 10 の正規分布に従う乱数列を 10 種類作る。サンプル数 n $=$ 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048 の乱数列を作る。
  - 標本から母平均の区間推定
    この 10 種類の乱数列のそれぞれから、母平均の区間推定を行え。信頼度 95%で、信頼区間 "$\pm 1.960 \frac{s}{\sqrt{n}}$" を用いよ。10 種類の結果をひとつの図にプロットせよ。 X 軸にサンプル数を Y 軸に平均値をとり、それぞれのサンプルから推定した平均とその信頼区間を示せ

box-muller 法による正規乱数生成

basic form: creates 2 normally distributed random variables, z_0 and z_1, from 2 uniformly distributed random variables, u_0 and u_1, in $(0, 1]$

$$z_0 = R\cos(\theta) = \sqrt{-2\ln u_0}\cos(2\pi u_1)$$
$$z_1 = R\sin(\theta) = \sqrt{-2\ln u_0}\sin(2\pi u_1)$$

polar form: 三角関数を使わない近似
u_0 and u_1: uniformly distributed random variables in $[-1, 1]$,
$s = u_0^2 + u_1^2$ (if $s = 0$ or $s \geq 1$, re-select u_0, u_1)

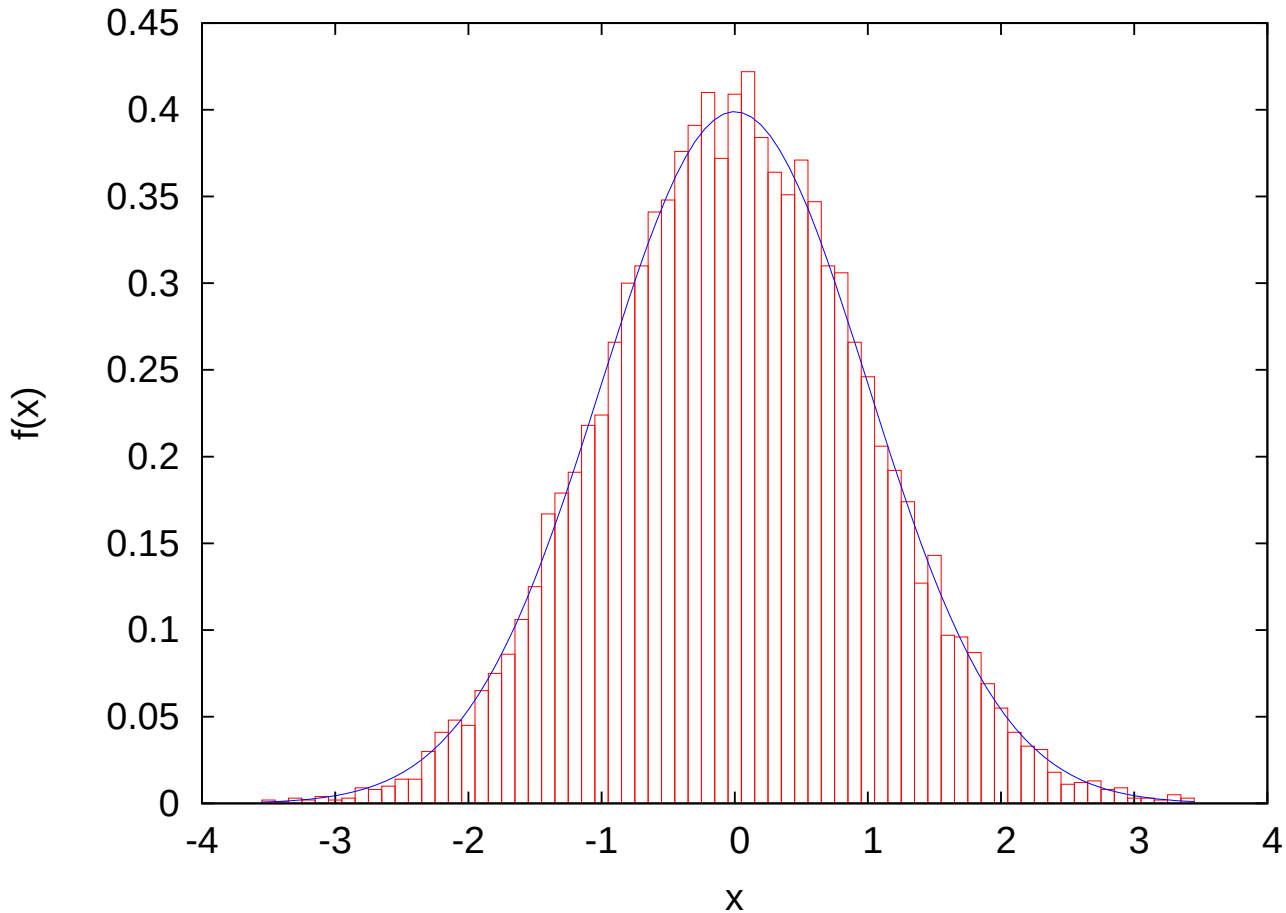
$$z_0 = u_0\sqrt{\frac{-2\ln s}{s}}$$
$$z_1 = u_1\sqrt{\frac{-2\ln s}{s}}$$

box-muller 法による正規乱数生成コード

```
# usage: box-muller.rb [n [m [s]]]
n = 1 # number of samples to output
mean = 0.0
stddev = 1.0

n = ARGV[0].to_i if ARGV.length >= 1
mean = ARGV[1].to_i if ARGV.length >= 2
stddev = ARGV[2].to_i if ARGV.length >= 3

# function box_muller implements the polar form of the box muller method,
# and returns 2 pseudo random numbers from standard normal distribution
def box_muller
  begin
    u1 = 2.0 * rand - 1.0  # uniformly distributed random numbers
    u2 = 2.0 * rand - 1.0  # ditto
    s = u1*u1 + u2*u2    # variance
  end while s == 0.0 || s >= 1.0
  w = Math.sqrt(-2.0 * Math.log(s) / s) # weight
  g1 = u1 * w  # normally distributed random number
  g2 = u2 * w  # ditto
  return g1, g2
end
# box_muller returns 2 random numbers.  so, use them for odd/even rounds
x = x2 = nil
n.times do
  if x2 == nil
    x, x2 = box_muller
  else
    x = x2
    x2 = nil
  end
  x = mean + x * stddev  # scale with mean and stddev
  printf "%.6f\n", x
end
```

正規乱数のヒストグラム作成

- 標準正規乱数のヒストグラムを作成し、正規分布であることを確認する
- 標準正規乱数を 10,000 個生成し、小数点 1 桁のビンでヒストグラムを作成



ヒストグラムの作成

- 少数点以下 1 桁でヒストグラムを作成する

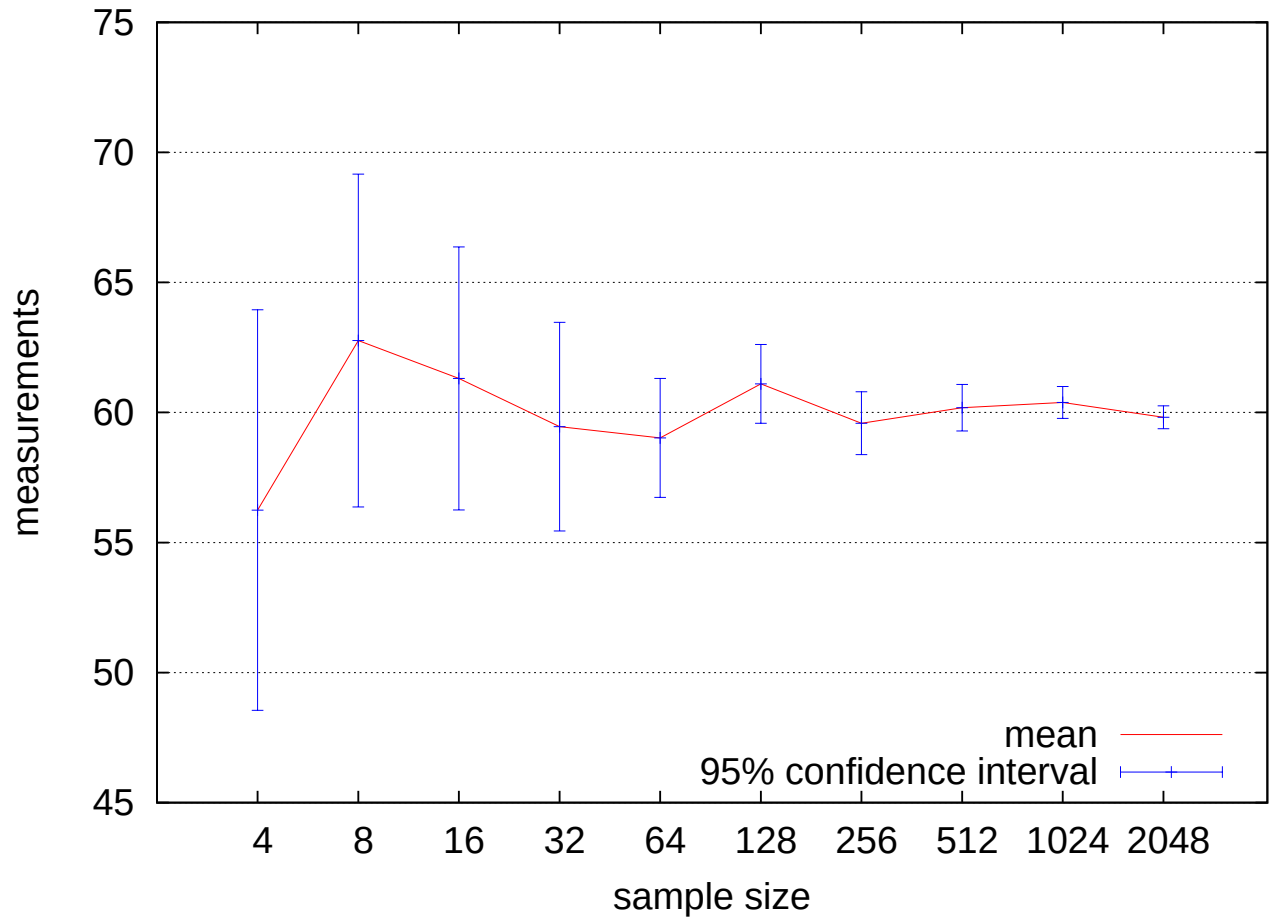
```
#
# create histogram: bins with 1 digit after the decimal point
#

re = /(-?\d*\.\d+)/ # regular expression for input numbers

bins = Hash.new(0)

ARGF.each_line do |line|
  if re.match(line)
    v = $1.to_f
    # round off to a value with 1 digit after the decimal point
    offset = 0.5    # for round off
    offset = -offset if v < 0.0
    v = Float(Integer(v * 10 + offset)) / 10
    bins[v] += 1 # increment the corresponding bin
  end
end

bins.sort{|a, b| a[0] <=> b[0]}.each do |key, value|
  puts "#{key} #{value}"
end
```

正規乱数のヒストグラムのプロット

```
set boxwidth 0.1
set xlabel "x"
set ylabel "f(x)"
plot "box-muller-histogram.txt" using 1:($2/1000) with boxes notitle, \
     1/sqrt(2*pi)*exp(-x**2/2) notitle with lines linetype 3
```

平均値の信頼区間とサンプル数の検証

サンプル数が増えるに従い、信頼区間は狭くなる



平均値の信頼区間のサンプル数による変化

課題 1: ホノルルマラソン完走時間のプロット

- ねらい: 実データから分布を調べる
- データ: 2012 年のホノルルマラソンの記録
  - http://results.sportstats.ca/res2012/honolulumarathon_m.htm
  - 完走者 24,070 人
- 提出項目
  1. 全完走者、男性完走者、女性完走者それぞれの、完走時間の平均、標準偏差、中間値
  2. それぞれの完走時間のヒストグラム
     - 3 つのヒストグラムを別々の図に書く
     - ビン幅は 10 分にする
     - 3 つのプロットは比較できるように目盛を合わせること
  3. それぞれの CDF プロット
     - ひとつの図に 3 つのプロットを書く
  4. オプション
     - 年代別や国別の CDF プロットなど自由
  5. 考察
     - データから読みとれることを記述
- 提出形式: レポートをひとつの PDF ファイルにして SFC-SFS から提出
- 提出〆切: 2013 年 5 月 16 日

ホノルルマラソンデータ

データフォーマット

Place	Chip Time	Pace /mi	#	Name	City	ST	CNT	Gender Plce/Tot	Category Plc/Tot	Category	@10km Split1	@21.1 Split2	@3
1	02:12:31	5:04	6	Kipsang, Wilson	Iten		KEN	1/12690	1/16	MElite	31:40	1:07:07	1:3
2	02:13:08	5:05	7	Geneti, Markos	Addis Ababa		ETH	2/12690	2/16	MElite	31:39	1:07:02	1:3
3	02:14:15	5:08	11	Kimutai, Kiplimo	Eldoret		KEN	3/12690	3/16	MElite	31:40	1:07:02	1:3
4	02:14:55	5:09	2	Ivuti, Patrick	Kangundo		KEN	4/12690	4/16	MElite	31:40	1:07:02	1:3
5	02:15:17	5:10	12	Arile, Julius	Kepenguria		KEN	5/12690	5/16	MElite	31:39	1:07:02	1:3
6	02:15:53	5:11	9	Bouramdane, Abderr	Champs De Cou		MAR	6/12690	6/16	MElite	31:40	1:07:01	1:3
7	02:18:27	5:17	8	Manza, Nicholas	Ngong Hills		KEN	7/12690	7/16	MElite	31:39	1:07:01	1:3
8	02:19:46	5:20	1	Chelimo, Nicholas	Ngong Hills		KEN	8/12690	8/16	MElite	31:40	1:07:02	1:3
9	02:25:23	5:33	20850	Harada, Taku	Nagoya-Shi	AI	JPN	9/12690	1/1238	M25-29	31:54	1:09:52	1:4
10	02:27:12	5:37	25474	Hagawa, Eiichi	Matsumoto	NA	JPN	10/12690	1/1501	M30-34	32:46	1:12:21	1:4

...

- Chip Time: 完走時間
- Category: MElite, WElite, M15-19, M20-24, ..., W15-29, W20-24, ...
  - "No Age"となっている人がいるので注意
- Country: 3-letter country code: e.g., JPN, USA
  - "UK"が交じっているので注意
- 完走者を抽出したら、総数が合っているかチェックすること

今回の演習: CCDF のプロット

twitter ユーザの following/follower の数の分布を CCDF にプロット

- データ: Kwak らによる 2009 年 7 月の twitter data
  - http://an.kaist.ac.kr/traces/WWW2010.html
  - twitter API を使って全ユーザ (4000 万以上) を crawl
    - API 変更、ユーザ増加によって、現在では出来ない
  - ここから、各ユーザの following/follower 数を抽出し、1 万人をサンプル

```
% head -10 twitter_degrees-10000.txt
# id followings followers
2058    1       1
11097   5       4
12329   375     1132
596043  63      97
638173  407     428
643423  2       18
659943  1958    1294
698823  503     344
730013  23      13
% ./make_ccdf.rb twitter_degrees-10000.txt > followings-ccdf.txt
% ./make_ccdf.rb -c 3 twitter_degrees-10000.txt > followers-ccdf.txt
```

今回の演習: CCDF のプロット (続き)

- CCDF 用データ形式: follow 数 人数 累積人数 CDF CCDF

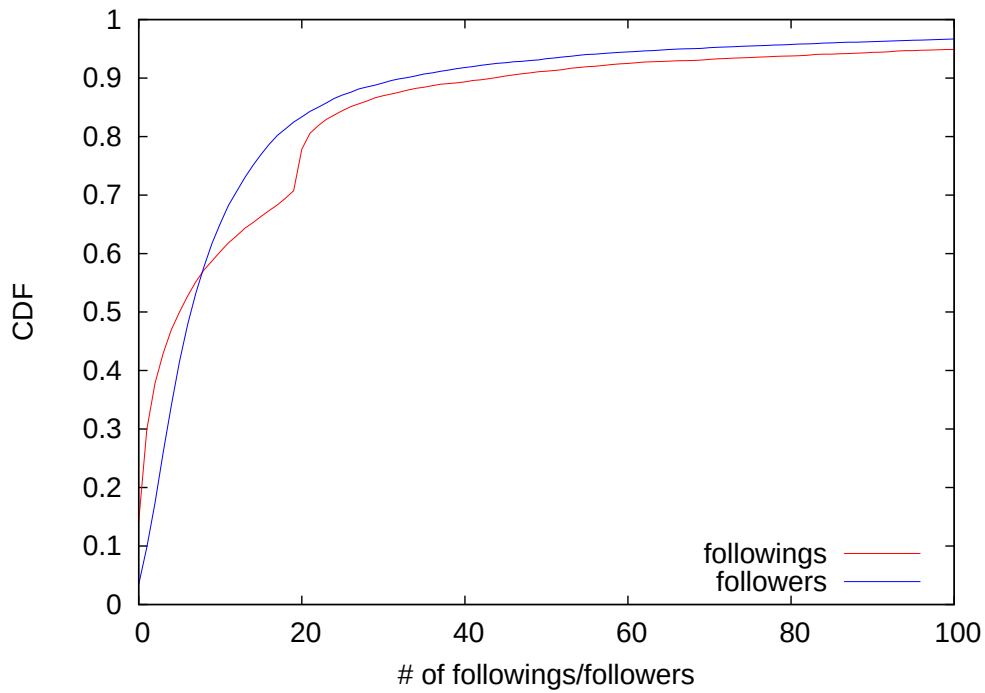
```
% cat followings-ccdf.txt
0       1407    1407    0.1407  1.0000
1       1586    2993    0.2993  0.8593
2       787     3780    0.3780  0.7007
3       513     4293    0.4293  0.6220
4       403     4696    0.4696  0.5707
5       302     4998    0.4998  0.5304
6       275     5273    0.5273  0.5002
7       242     5515    0.5515  0.4727
8       202     5717    0.5717  0.4485
9       158     5875    0.5875  0.4283
...
2134    1       9991    0.9991  0.0010
2167    1       9992    0.9992  0.0009
2510    1       9993    0.9993  0.0008
2512    1       9994    0.9994  0.0007
2657    1       9995    0.9995  0.0006
3410    1       9996    0.9996  0.0005
5042    1       9997    0.9997  0.0004
6605    1       9998    0.9998  0.0003
11350   1       9999    0.9999  0.0002
12335   1       10000   1.0000  0.0001
```

CCDF を集計するスクリプト: make_ccdf.rb

```
#!/usr/bin/env ruby
# create ccdf: usage: make_ccdf.rb [-c n] file
#
require 'optparse'

re = /^(\d+)\s+(\d+)\s+(\d+)/
col_no = 2 # column no (1 origin)

OptionParser.new {|opt|
  opt.on('-c VAL', Integer) {|v| col_no = v}
  opt.parse!(ARGV)
}

counts = Hash.new(0)
n = 0
ARGF.each_line do |line|
  if re.match(line)
    val = $~[col_no].to_i
    counts[val] += 1
    n += 1
  end
end

cum = 0
cdf = 0.0
total = counts.length
counts.sort.each do |key, value|
  comp = 1.0 - cdf
  cum += value
  cdf = Float(cum) / n
  printf "%d\t%d\t%d\t%.4f\t%.4f\n", key, value, cum, cdf, comp
end
```

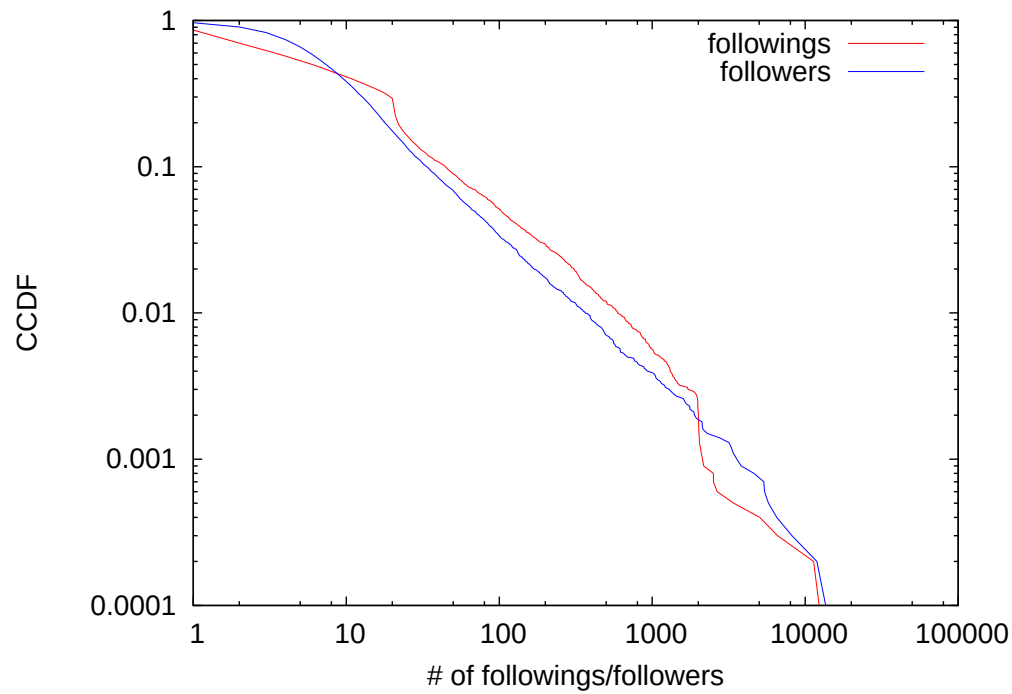
CDF の表示: gnuplot スクリプト

```
set xlabel "# of followings/followers"
set ylabel "CDF"
set key bottom
set xrange [0:100]

plot "followings-ccdf.txt" using 1:4 title 'followings' with lines, \
"followers-ccdf.txt" using 1:4 title 'followers' with lines lt 3
```



CCDF の表示: gnuplot スクリプト

```
set logscale
set xlabel "# of followings/followers"
set ylabel "CCDF"

plot "followings-ccdf.txt" using 1:5 title 'followings' with lines, \
"followers-ccdf.txt" using 1:5 title 'followers' with lines lt 3
```



演習の考察

CDF/CCDF から分かる事

- フォロー数、フォロワー数ともにべき分布に従う
  - 極端に少ない/多い部分はその限りでない
- フォロー数は、20 と 2000 あたりにギャップ
  - 20: 初期設定でフォローする 20 人を薦められる
  - 2000: 以前は最大 2000 人までしかフォローできなかった

まとめ

第 5 回 多様性と複雑さ

- ロングテール
- Web アクセスとコンテンツ分布
- べき乗則と複雑系
- 演習: べき乗則解析

次回予定

第 6 回 相関 (5/15)

- オンラインお勧めシステム
- 距離とエントロピー
- 相関係数
- 演習: 相関

5/22 休講

参考文献

[1] Ruby official site. `http://www.ruby-lang.org/`

[2] gnuplot official site. `http://gnuplot.info/`

[3] Mark Crovella and Balachander Krishnamurthy. *Internet measurement: infrastructure, traffic, and applications*. Wiley, 2006.

[4] Pang-Ning Tan, Michael Steinbach and Vipin Kumar. *Introduction to Data Mining*. Addison Wesley, 2006.

[5] Raj Jain. *The art of computer systems performance analysis*. Wiley, 1991.

[6] Toby Segaran. (當山仁健 鴨澤眞夫 訳). 集合知プログラミング. オライリージャパン. 2008.

[7] Chris Sanders. (高橋基信 宮本久仁男 監訳 岡真由美 訳). 実践パケット解析 第 *2* 版 — *Wireshark* を使ったトラブルシューティング. オライリージャパン. 2012.

[8] あきみち、空閑洋平. インターネットのカタチ. オーム社. 2011.

[9] 井上洋, 野澤昌弘. 例題で学ぶ統計的方法. 創成社, 2010.

[10] 平岡和幸, 掘玄. プログラミングのための確率統計. オーム社, 2009.