

インターネット計測とデータ解析 第2回

長 健二郎

2011年5月18日

前回のおさらい

本授業のテーマ

- ▶ いろいろな切口からインターネットの実態を考える
 - ▶ 容易に計測できないものをどう計るか
 - ▶ 大量データからいかに情報を抽出する

第1回 イン트로ダクション

- ▶ ネットワーク計測とインターネット計測
- ▶ ネットワーク管理ツール
- ▶ 計測ツール
- ▶ 演習:ruby 入門

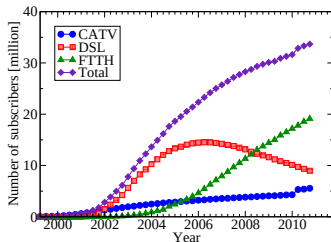
今日のテーマ

インターネットのサイズを計る

- ▶ ユーザ数、ホスト数
- ▶ ウェブページ数
- ▶ 精度 誤差 有効数字
- ▶ グラフによる可視化
- ▶ 演習:gnuplot によるグラフ描画

インターネットのユーザ数 (日本)

- ▶ 総務省 通信利用動向調査
 - ▶ 9408 万人 人口普及率 78.0% (2009 年末)
 - ▶ 無作為抽出アンケート方式
 - ▶ 地域及び都市規模を層化基準とした層化二段抽出
 - ▶ 世帯調査 サンプル数 6,256 世帯 有効回答数 4,547
 - ▶ ちなみに全国世帯数 5336 万 (2010/03)
- ▶ 総務省 ブロードバンド契約数
 - ▶ 電気通信事業者からの報告
 - ▶ 契約数 3402 万 (2010/09)

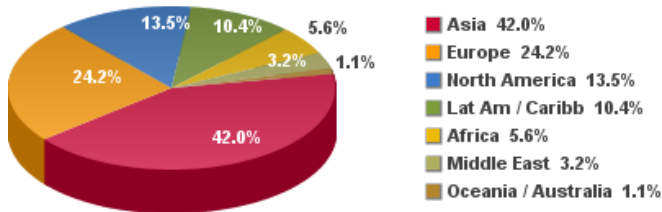


出典: 総務省 ブロードバンド契約数の推移

世界のインターネットユーザ数

- ▶ 世界 19.7 億人 人口比普及率 28.7% (2010/06)

Internet Users in the World Distribution by World Regions - 2010



Source: Internet World Stats - www.internetworldstats.com/stats.htm

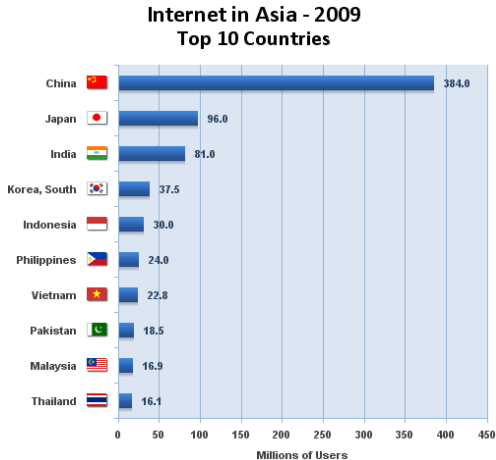
Basis: 1,966,514,816 Internet users on June 30, 2010

Copyright © 2010, Miniwatts Marketing Group

出典: Internet World Stats <http://www.internetworldstats.com/>

アジアのインターネットユーザ数

- ▶ 中国が圧倒的 4.2 億人 人口比普及率 31.6% (2010/06)



Source: www.internetworldstats.com/stats3.htm
Estimated Internet users in Asia 764,435,900 for 2009
Copyright © 2010, Miniwatts Marketing Group

出典: Internet World Stats <http://www.internetworldstats.com/>

インターネットに繋がっている端末数

インターネットに繋がっているという定義は？

- ▶ なんらかの形でインターネット上のデータにアクセスできる
 - ▶ web が見られる
 - ▶ 電子メールが届く
 - ▶ 技術的にはかるのは難しいが、
 - ▶ 2010 年 世界の携帯電話契約数: 50 億
 - ▶ 米 IDC 社調査 2009 年 世界 PC 出荷台数 約 3 億台
- ▶ IP プロトコルで通信できる (NAT の裏側の端末を含む)
- ▶ グローバル IP アドレスを持つ (双方向で IP 通信可能)

ホスト数をはかる

目的

- ▶ インターネットに繋がっているコンピュータ数の把握
 - ▶ NAT の普及で困難
- ▶ IP アドレス利用状況を把握する
 - ▶ IP アドレスは限られた資源
 - ▶ 割り当て (回収) ポリシーへの反映
 - ▶ IPv4 アドレスの枯渇問題

方法

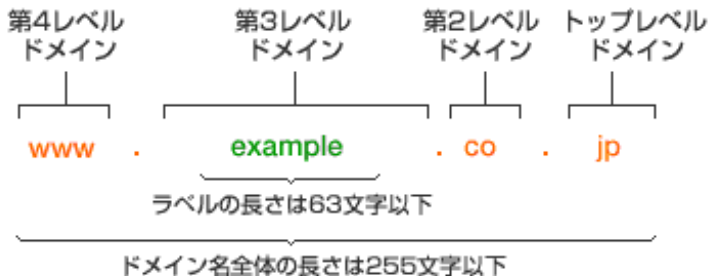
- ▶ DNS から網羅的に調べる
- ▶ IP アドレス空間 (2^{32}) を網羅的に調べる
- ▶ サンプルングして推測
 - ▶ アドレスブロックの利用形態の違いから容易ではない

Domain Name System(DNS) の仕組み (1/3)

JPNIC 「ドメイン名のしくみ」より

▶ <http://www.nic.ad.jp/ja/dom/system.html>

ドメイン名の構成

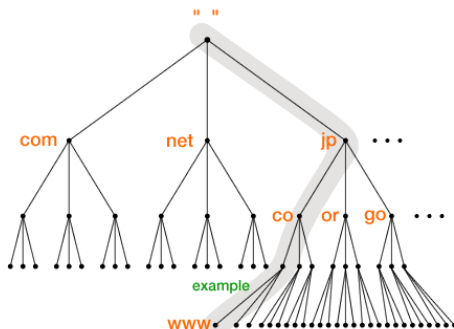


出典:JPNIC

Domain Name System(DNS) の仕組み (2/3)

DNS の構造

- ▶ root を頂点としたツリー構造
- ▶ 各ドメインには「ネームサーバー」がいてデータベースを分散管理
 - ▶ 配下のドメイン名と IP アドレスの関係を管理
 - ▶ 下位ドメインのネームサーバーへの参照管理

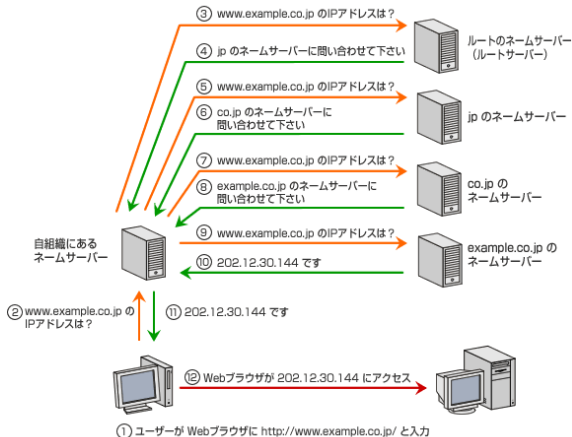


ドメイン名空間 出典:JPNIC

Domain Name System(DNS) の仕組み (3/3)

DNS における名前解決の方法

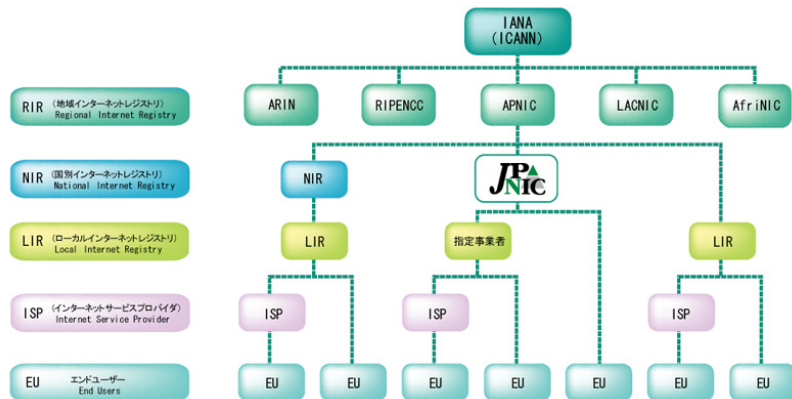
- ▶ 名前解決: ドメイン名を対応する IP アドレスに変換
 - ▶ 逆引き: IP アドレスをドメイン名に変換 (逆引きツリー)



名前解決の流れ (`www.example.co.jp` の例) 出典: JPNIC

IP アドレスの割り当て管理

▶ IANA → RIR → NIR → LIR

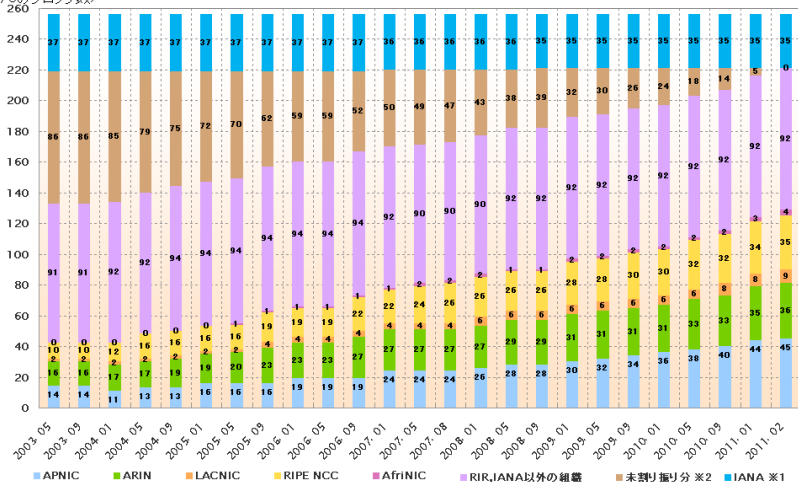


IP アドレス管理の階層構造 出典:JPNIC

IPv4 アドレス在庫の枯渇

- ▶ 2011/2/1 IANA の在庫が枯渇
- ▶ 2011/4/15 APNIC と JPNIC の在庫が枯渇

(/8のブロック数)

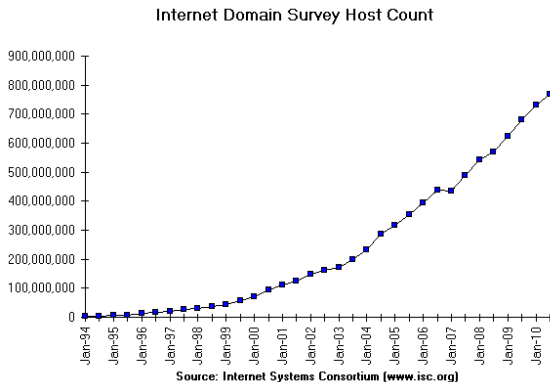


地域インターネットレジストリ (RIR) ごとの IPv4 アドレス配分状況 出典:JPNIC

古典的なホスト数推計方法

The ISC Domain Survey (DNS による調査)

▶ 6.8 億ホスト (2009/07)



出典: ISC domain survey <http://www.isc.org/solutions/servey>

The ISC Domain Survey

計測方法

- ▶ 1987-1997:DNS に登録されたホスト数をカウント (RFC1296)
 - ▶ DNS の委譲ツリーを辿って、各ゾーンからゾーンデータ転送を試みる
 - ▶ ゾーンデータ中の「A レコード」を数える
 - ▶ ゾーンデータ転送を許可しない分を補正するため、ゾーン転送の成功率を使う
- ▶ 1998-:DNS に登録されたユニークな IP アドレス数をカウント
 - ▶ 逆引きの委譲ツリーを辿って、存在する /24 を見つける
 - ▶ 見つかった /24 の全ての IP アドレス (1-254) を逆引きし「PTR レコード」の登録があるか調べる
 - ▶ PTR レコードがあるが存在しないホストがあるため、発見したアドレスの 1%をランダムサンプリングして ping、成功率を補正に使う

制約

- ▶ DNS に登録されていないものはカウントされない
- ▶ DNS に登録だけされて存在しないホストの補正精度
- ▶ NAT の背後にいるホスト数はカウントできない

IP アドレス空間の網羅的調査

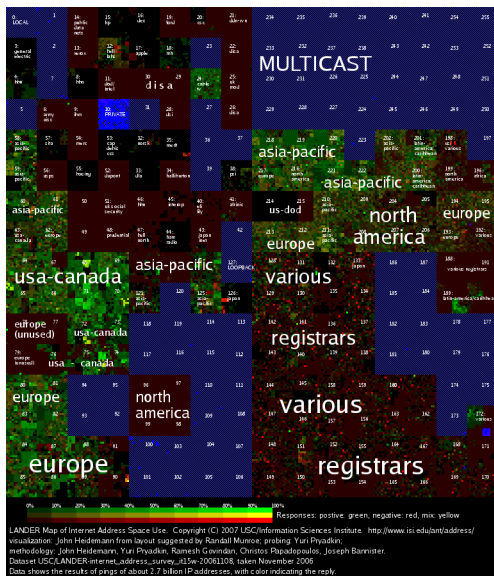
- ▶ Heidemann らの 2006/11 の計測データ
- ▶ 割り当て済みの IP アドレス全てに ping
- ▶ 調べたアドレスの 93%は応答なし (firewall, etc)

address type	number	% of addrs	% of probed
IPv4 addresses	4,290M	100%	
reserved	1,160M	27%	
allocated	3,140M	73%	
unprobed (mcast, etc)	342M	8%	
probed	2,800M	65%	100%
replies	187M	4.4%	6.7%
positive replies	103M	3.6%	3.7%
negative replies	84M	2.0%	3.0%
non-replies	2,610M	61%	93%

J. Heidemann, Y. Pradkin, R. Govindan, C. Papadopoulos, G. Bartlett, J. Bannister.
Census and survey of the visible internet.

ACM IMC'08. pp169-182. Vouliagmeni, Greece. October 2008.

IP アドレス空間の利用状況の可視化

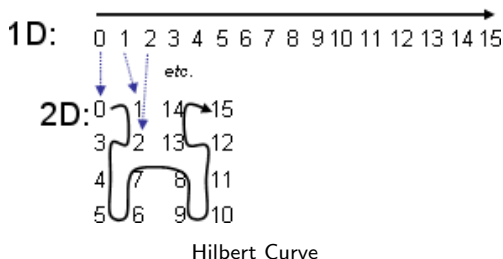


<http://www.isi.edu/ant/address/>

IP アドレス空間の利用状況の可視化 (つづき)

可視化手法

- ▶ Hilbert Curve による空間表現 (連続空間が隣接する、再帰的)
- ▶ 各点は/16 ブロック (64k addrs) の平均
- ▶ positive:green, negative:red, mix:yellow



ウェブページ数をはかる

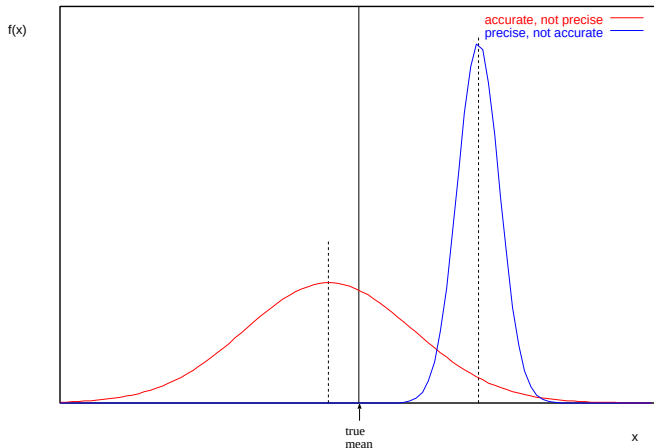
- ▶ ウェブページの定義? 動的ページ (calendar, etc) が増加
- ▶ crawling robot によりデータ収集可能
 - ▶ 人気サイトから始めてリンクを辿る
- ▶ 大規模検索システムはある程度情報を持っている、公開はされていない
- ▶ netcraft: web server survey 227 million sites in 2010/09
- ▶ google: indexed 1 trillion (10^{12}) unique URLs in 2008
 - ▶ <http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>

正確度と精度、誤差

正確度 (accuracy): 測定値と真値とのずれ

精度 (precision): 測定値のばらつきの幅

誤差 (error): 真値からのずれ、その不確かさの範囲



いろいろな誤差

測定誤差

- ▶ 系統誤差 (条件を把握できれば補正可能)
 - ▶ 器械的誤差、理論的誤差、個人的誤差
- ▶ 偶然誤差 (ノイズ、観測を繰り返せば精度向上)

計算誤差

- ▶ まるめ誤差
- ▶ 打ち切り誤差
- ▶ 情報落ち
- ▶ 桁落ち
- ▶ 誤差の伝搬

サンプリング誤差

- ▶ 標本調査を行う場合、普通は真値は不明
- ▶ 標本誤差: 真値との差の確率的なばらつきの幅

有効数字と有効桁数

1.23 の有効数字は 3 桁 ($1.225 \leq 1.23 < 1.235$)
表記

表記	有効桁数	
12.3	3	
12.300	5	
0.0034	2	
1200	4	(あいまい、 1.200×10^3)
2.34×10^4	3	

計算

- ▶ 計算途中は桁数が大きいまま計算
 - ▶ 筆算などの場合は 1 桁多く取ればよい
- ▶ 最終的な数字に有効桁数を適用

基本ルール

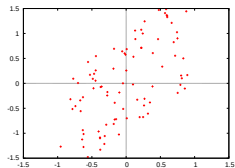
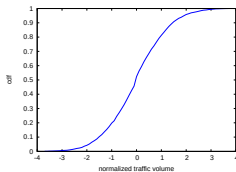
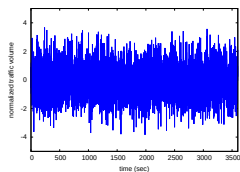
- ▶ 加減算: 桁数が少ないものに合わせる
 - ▶ $1.23 + 5.724 = 6.954 \Rightarrow 6.95$
- ▶ 乗除算: もとの有効数字が最も少ないものに合わせる
 - ▶ $4.23 \times 0.38 = 1.6074 \Rightarrow 1.6$

コンピュータの計算精度

- ▶ integer (32/64bits)
 - ▶ 32bit signed integer (2G までしかカウントできない)
- ▶ 32bit floating point (IEEE 754 single precision): 有効桁数 7
 - ▶ sign:1bit, exponent:8bits, mantissa:23bits
 - ▶ $16,000,000 + 1 = 16,000,000!!$
- ▶ 64bit floating point (IEEE 754 double precision): 有効桁数 15
 - ▶ sign:1bit, exponent:11bits, mantissa:52bits

グラフ描画

直観的にデータの性質を把握するには、いくつかの統計的手法を用いてグラフを描画してみる



グラフ描画のガイドライン

読み手の立場にたって、分かり易いグラフを描画する

- ▶ XY 軸のラベルを明確に
- ▶ XY 軸の目盛りと単位を明確に
- ▶ 個々の直線曲線にもラベルを付ける
- ▶ 適切なフォントとサイズを使う
- ▶ 慣習に従う: 0 を起点にする、数学シンボルや略称の使用など
- ▶ ばらつきを示す (平均値だけでは不十分)
- ▶ グラフの範囲を適切か
- ▶ ひとつのグラフで多くを示さない
- ▶ 異なるデータを比較する場合は、適切な正規化を行う
- ▶ グラフ同士を比較する場合は、XY 軸のスケールを合わせる
- ▶ 技術系は円グラフや 3D 効果グラフは使わない
- ▶ 色を使う場合
 - ▶ 白黒印刷しても読めるように配慮
 - ▶ プロジェクタ投影も配慮 (例:黄色は避ける)

データ変数

- ▶ 一変数解析 (univariate analysis)
 - ▶ 変数をひとつずつ独立して扱う
- ▶ 多変量解析 (multivariate analysis)
 - ▶ 複数の変数を同時に扱う

生データのグラフ化

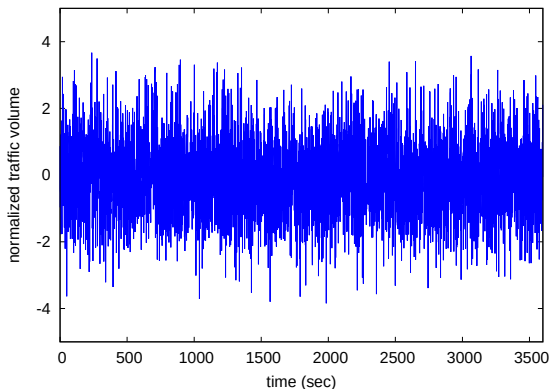
代表的なグラフ

- ▶ 時系列グラフ
- ▶ ヒストグラム
- ▶ 確率グラフ
- ▶ 散布図

時系列グラフ

変数の時間変化を見る

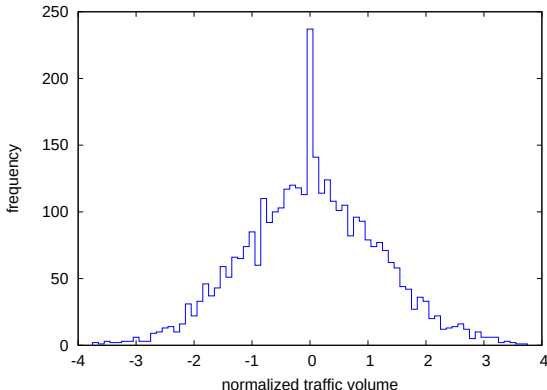
- ▶ X 軸に時間、Y 軸に変数値
- ▶ 時系列グラフから分かること
 - ▶ 位置の変化
 - ▶ ばらつきの変化
 - ▶ 外れ値の存在



ヒストグラム (1/2)

変数の分布の仕方を見る

- ▶ データを同じ幅のビンに分ける
- ▶ 各ビンのデータ数を数える
- ▶ X 軸:ビンの値 Y 軸:データ数



ヒストグラム (2/2)

ヒストグラムから分かる事

- ▶ 分布の中心 (位置)
- ▶ 分布の広がり
- ▶ 分布の偏り
- ▶ 外れ値の存在
- ▶ 複数のモードの存在 (山が複数あるか)

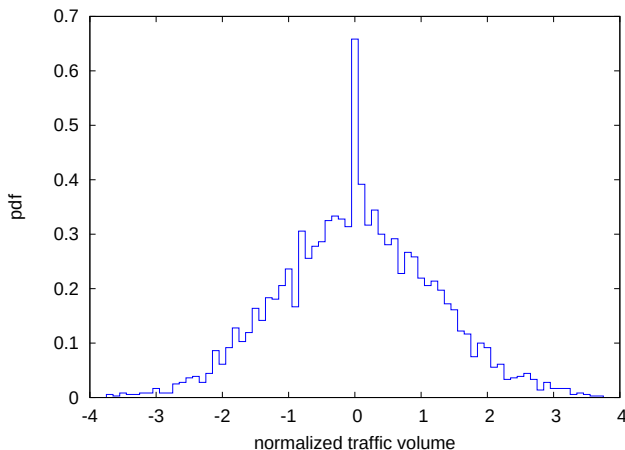
ヒストグラムの制約

- ▶ 適切なビン幅を選ぶ必要
 - ▶ 小さ過ぎると各ビンのサンプル数が足りなくなる
 - ▶ 大き過ぎると分布の詳細が分からない
 - ▶ 偏りの大きい分布では適切なビン幅の選択は難しい
- ▶ 十分なサンプル数が必要

確率密度関数 (probability density function; pdf)

- ▶ 合計面積が1となるように出現数を正規化
 - ▶ 出現数を (総データ数 × ビン幅) で割る
- ▶ 確率密度関数: 確率変数 X が x という値をとる確率

$$f(x) = P[X = x]$$



累積分布関数 (cumulative distribution function; cdf)

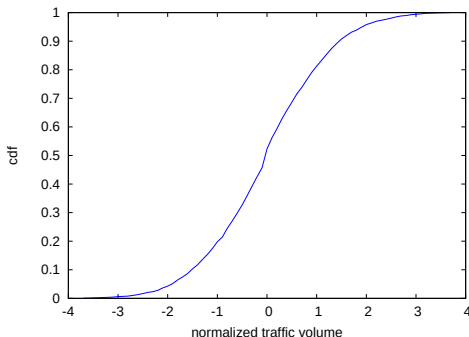
- ▶ 密度関数: x をいう値を観測する確率

$$f(x) = P[X = x]$$

- ▶ 累積分布関数: x 以下の値を観測する確率

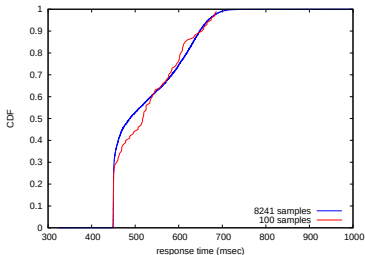
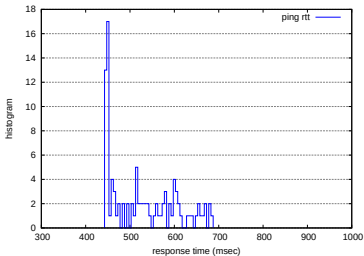
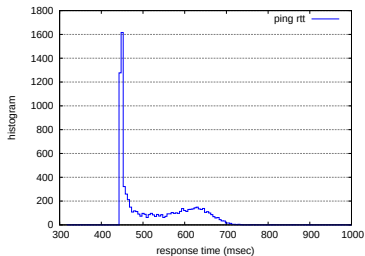
$$F(x) = P[X \leq x]$$

- ▶ 分布の偏りが大きい、サンプル数が少ない、外れ値が無視できない場合などは、ヒストグラムより有効



ヒストグラムと CDF の比較

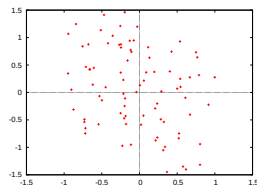
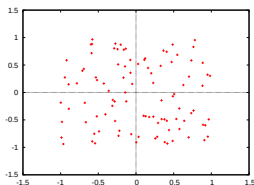
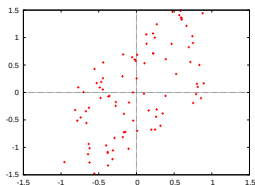
- ▶ CDF の場合、ビン幅やサンプル数不足を考慮しなくていい



(左) 元データ (右)100 サンプル (下)CDF

散布図 (scatter plots)

- ▶ 2 つの変数の関係を見るのに有効
 - ▶ X 軸: 変数 X
 - ▶ Y 軸: それに対応する変数 Y の値
- ▶ 散布図で分かる事
 - ▶ X と Y に関連があるか
 - ▶ 無相関、正の相関、負の相関
 - ▶ 外れ値の存在があるか



例: (左) 正の相関 0.7 (中) 無相関 0.0 (右) 負の相関 -0.5

グラフ描画ツール

- ▶ gnuplot
 - ▶ コマンドラインツール、スクリプトで自動化し易い
 - ▶ <http://gnuplot.info/>
- ▶ grace
 - ▶ 使い易い GUI
 - ▶ 細かい仕上げ調整が可能
 - ▶ <http://plasma-gate.weizmann.ac.il/Grace/>

演習: gnuplot

- ▶ gnuplot を使って簡単なグラフを書く
- ▶ 例題: P. K. Janert "Gnuplot in Action" から転載
 - ▶ <http://web.sfc.keio.ac.jp/~kjc/classes/sfc2011s-measurement/marathon.txt>
 - ▶ <http://web.sfc.keio.ac.jp/~kjc/classes/sfc2011s-measurement/prices.txt>

ヒストグラム

▶ ある市民マラソンの完走タイムの分布

```
plot "marathon.txt" using 1:2 with boxes
```

グラフを見やすくする (右側)

```
set title "marathon finish time distribution"
```

```
set boxwidth 1
```

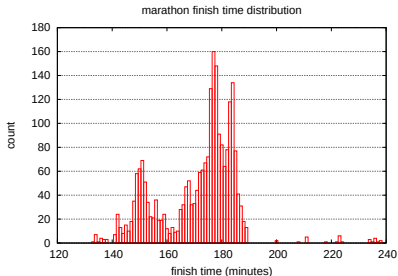
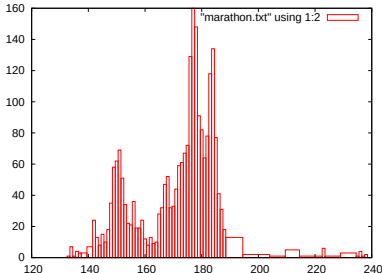
```
set xlabel "finish time (minutes)"
```

```
set ylabel "count"
```

```
set yrange [0:180]
```

```
set grid y
```

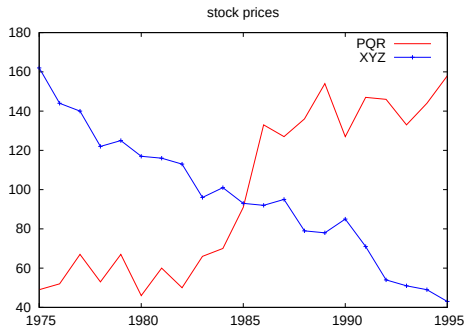
```
plot "marathon.txt" using 1:2 with boxes notitle
```



時系列プロット

▶ PQR 社と XYZ 社の株価推移

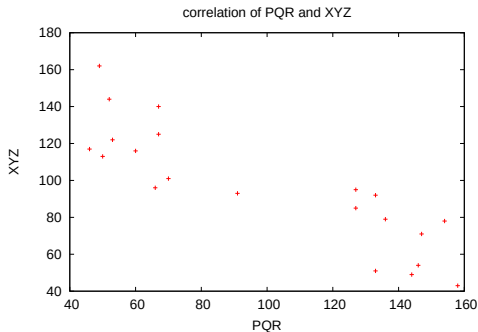
```
set title "stock prices"  
plot "prices.txt" using 1:2 title "PQR" with lines, \  
"prices.txt" using 1:3 title "XYZ" with linespoints linetype 3
```



散布図

▶ PQR 社と XYZ 社の株価の相関

```
set title "correlation of PQR and XYZ"  
set xlabel "PQR"  
set ylabel "XYZ"  
plot "prices.txt" using 2:3 notitle with points
```



まとめ

インターネットのサイズを計る

- ▶ ユーザ数、ホスト数
- ▶ ウェブページ数
- ▶ 精度 誤差 有効数字
- ▶ グラフによる可視化
- ▶ 演習:gnuplot によるグラフ描画

次回予定

第3回 データの記録とログ解析 (5/25)

- ▶ データフォーマット
- ▶ ログ解析手法
- ▶ 演習:ログデータと正規表現