インターネット計測とデータ解析 第7回

長 健二朗

2011年6月22日

前回のおさらい

インターネットの特徴量を計る

- 遅延、パケットロス、ジッタ
- フロー計測
- 相関と多変量解析
- 主成分分析
- 演習:多変量と相関

今日のテーマ

インターネットの多様性と複雑さを計る

- サンプリング
- 統計解析 (ヒストグラム、大数の法則)
- 演習:ヒストグラム、CDF
- **課題 2**

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
- 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

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)

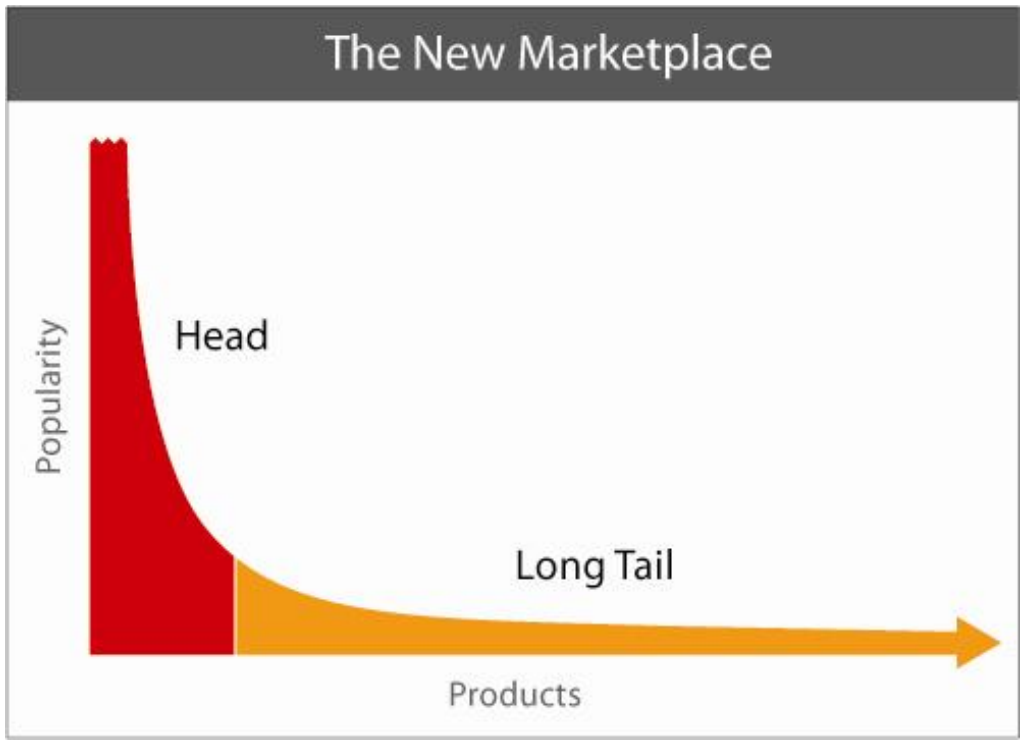
- 自己相似性
- 長期依存性

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/

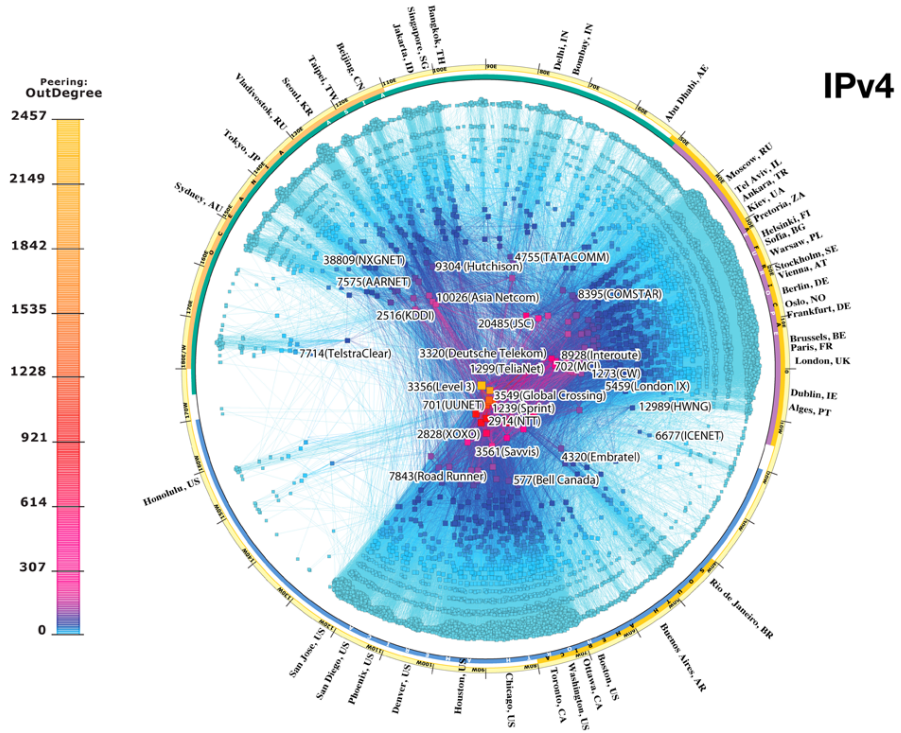
インターネットの AS 構造の例

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

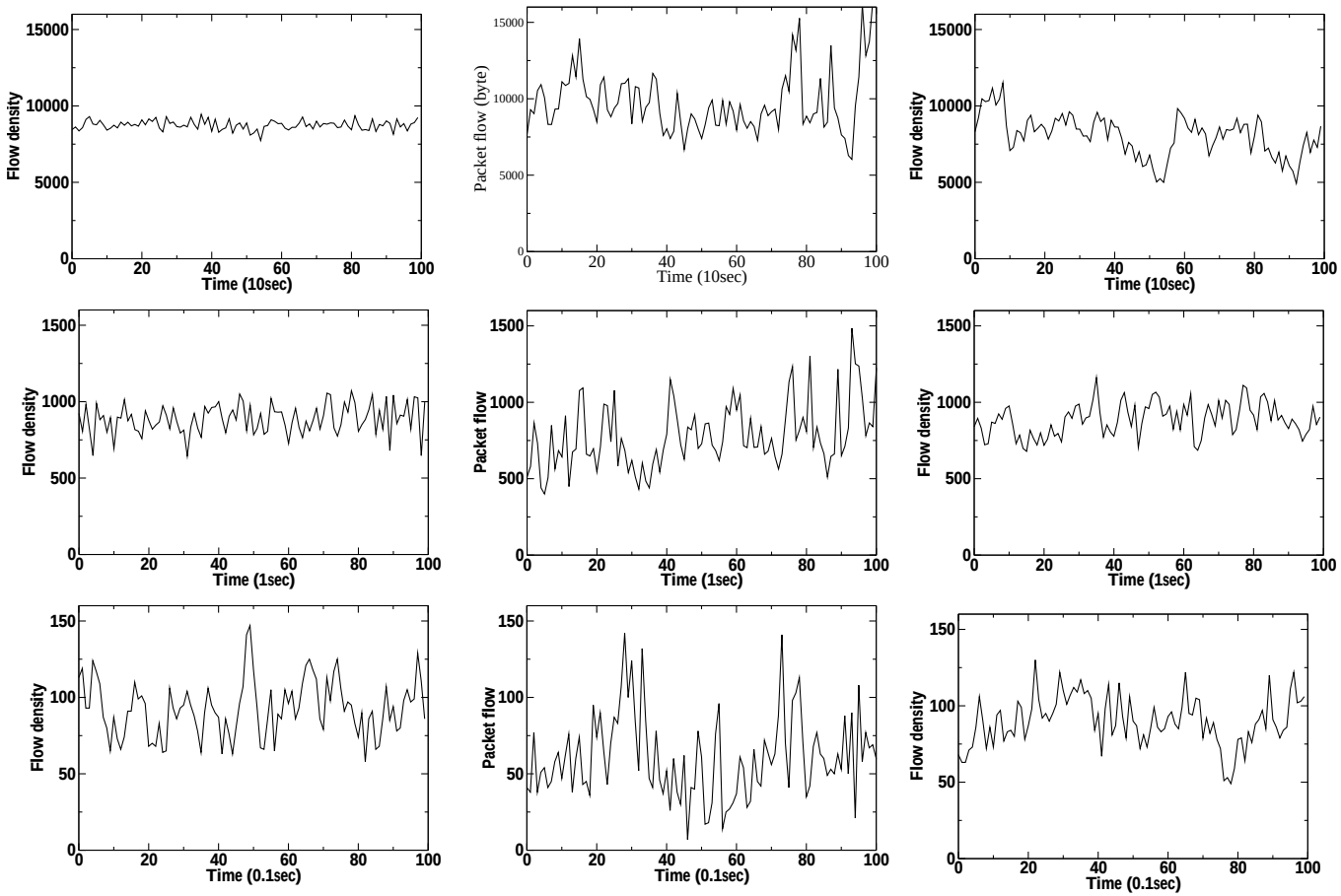
AS-level INTERNET GRAPH



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



インターネットのデータの多様性

- 観測する場所によって異なる挙動が見える
  - 国、地域、時間
  - 企業と大学と家庭、バックボーンとアクセスネットワーク

典型的なネットワークは存在しない

- 多様性をどうやって計るか、表すか
- どうサンプリングするか

サンプリング

- 全数調査: ほとんどの場合は非現実的
- サンプリングが必要になる

インターネット計測におけるサンプリング

- 測定場所
- 時間、期間
- パケット、フロー

パケットのサンプリング方法

- カウンタベースの 1/N サンプリング (決定論的)
  - 実装が簡単、広く使われている
  - 測定対象と同期してしまう可能性
- 確率的 1/N サンプリング
  - パケットごとにサイコロを振って決める
- 時間によるサンプリング
  - 例: 毎時最初の 1 分を計測
- フローベースのサンプリング
  - 新しいフローは確率的にサンプル
  - 選んだフローのパケットは全部測定
  - フローの挙動解析が可能
- 他にも様々な方法が存在

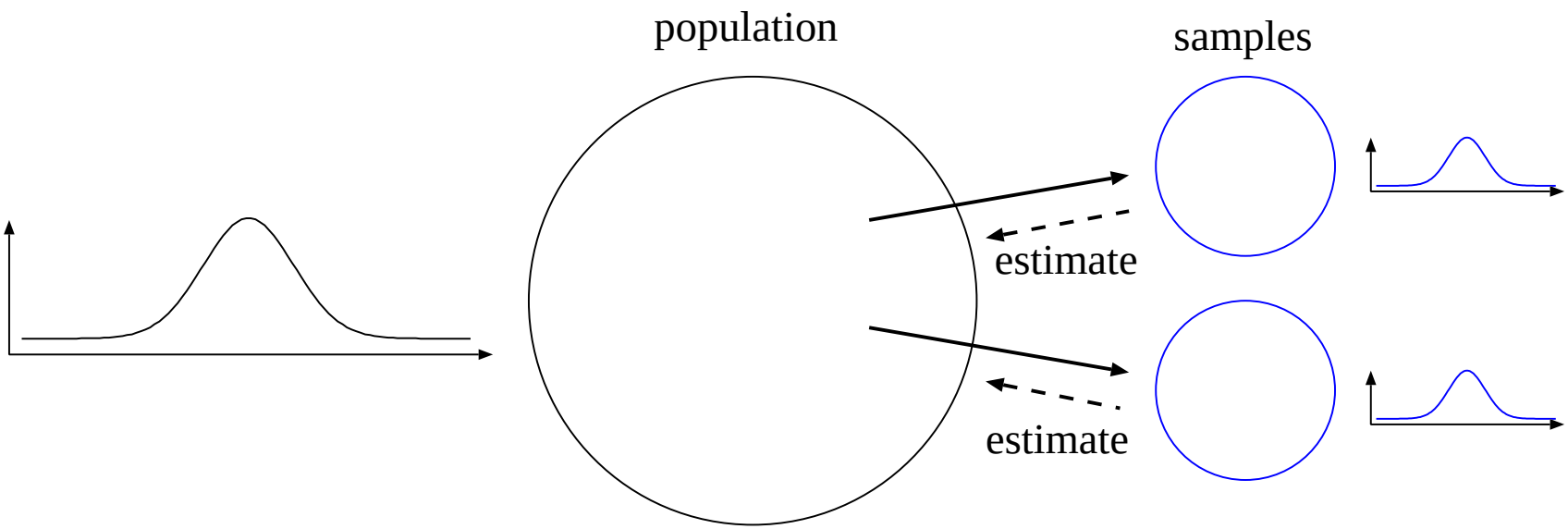
サンプリング: 標本と母集団

要約と推測

- 要約統計量 (平均、標準偏差など) は分布の特徴を要約して表す数値
- 推測統計は標本 (サンプル) から母集団の性質を統計的に推測する

母集団 (population): 全体のデータ、多くの場合入手不可能

- 標本 (sample) から母集団の性質を推定する必要
- 変数: 母集団の特徴 (固定)
- 統計: 標本からの推定値 (ゆらぎを持つ変数)



期待値

確率変数 X の期待値 $E(X)$ (平均を表す)

- 離散型

$$E(X) = \mu = \sum_{i=1}^{n} x_i p_i$$

- 連続型

$$E(X) = \mu = \int_{-\infty}^{\infty} x f(x) dx$$

期待値の性質

- $E(c) = c$
- $E(X + c) = E(X) + c$
- $E(cX) = cE(X)$
- $E(X + Y) = E(X) + E(Y)$

標本平均

- 標本平均 (sample mean): $\bar{x}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 標本分散 (sample variance): s^2

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- 標本標準偏差 (sample standard deviation): s
- 注: 二乗和を n ではなく $(n-1)$ で割る
  - 自由度 (degree of freedom): 二乗和の独立変数は $\bar{x}$ があるため 1 減る

大数の法則と中心極限定理

大数の法則

- サンプル数が増えるに従い標本平均は母平均に近付く

中心極限定理

- 元の分布に関わらず (十分なサンプル数があれば) 標本平均は近似的に正規分布に従う $N(\mu, \sigma^2/n)$
- 母集団が正規分布の場合は、n が小さくてもこの関係が成立する

標準誤差 (standard error)

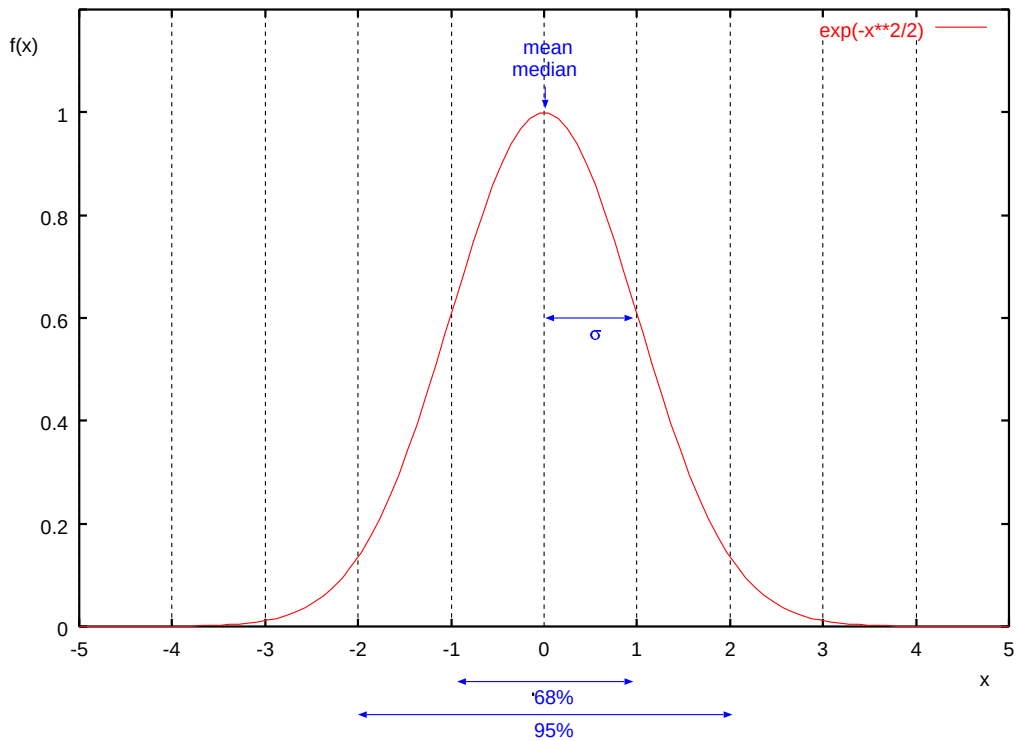
標準誤差: 標本平均の標準偏差 (SE)

$$SE = \sigma/\sqrt{n}$$

- サンプル数 n を増やすと精度が改善
  - 標準誤差は $1/\sqrt{n}$ に (ゆっくり) 減少
- 正規母集団 $N(\mu, \sigma)$ から取った標本平均の分布は平均 μ 標準偏差 $\mathrm{SE} = \sigma/\sqrt{n}$ の正規分布となる

正規分布 (normal distribution)

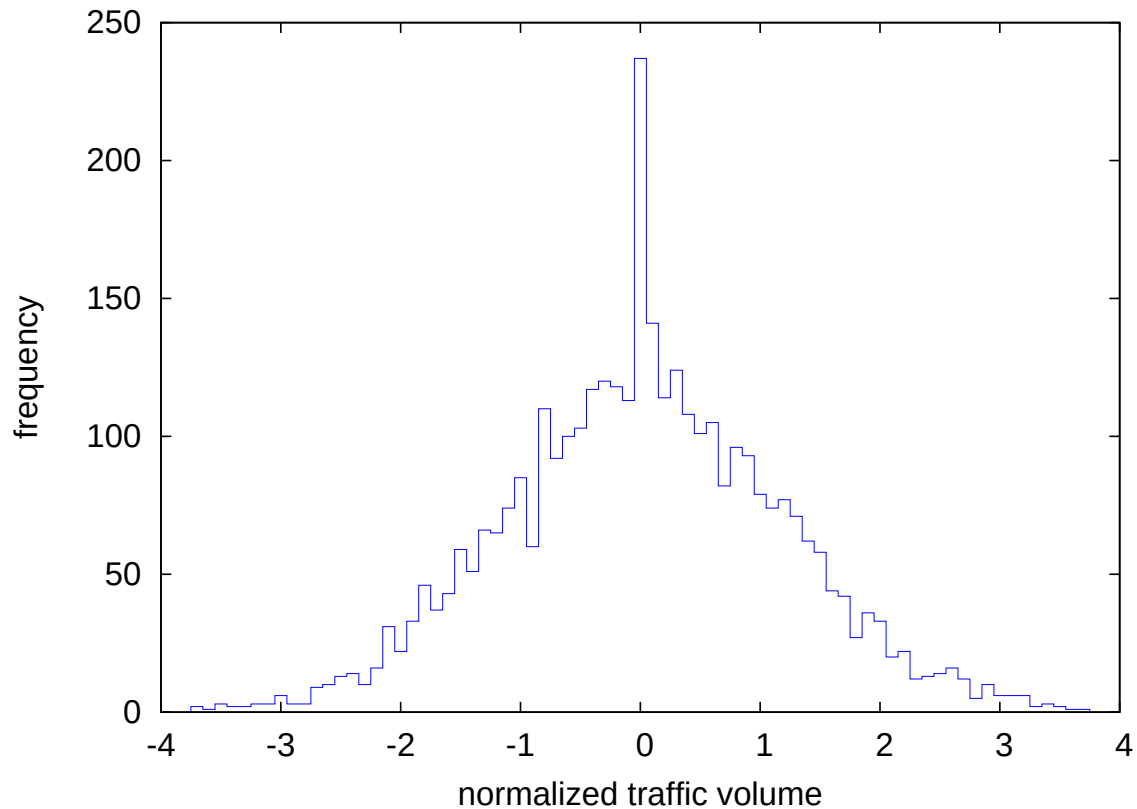
- つりがね型の分布、ガウス分布とも呼ばれる
- 2 つの変数で定義: 平均 μ、分散 σ^2
- 乱数の和は正規分布に従う
- 標準正規分布: $\mu = 0, \sigma = 1$
- 正規分布ではデータの
  - 68%は $(mean \pm stddev)$
  - 95%は $(mean \pm 2stddev)$ の範囲に入る



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

ビン幅の決め方

スタージェスの方法: ビン数 k、サンプル数 n

$$k = \log_2 n + 1$$

スコットの方法: ビン幅 h、標準偏差 σ、サンプル数 n

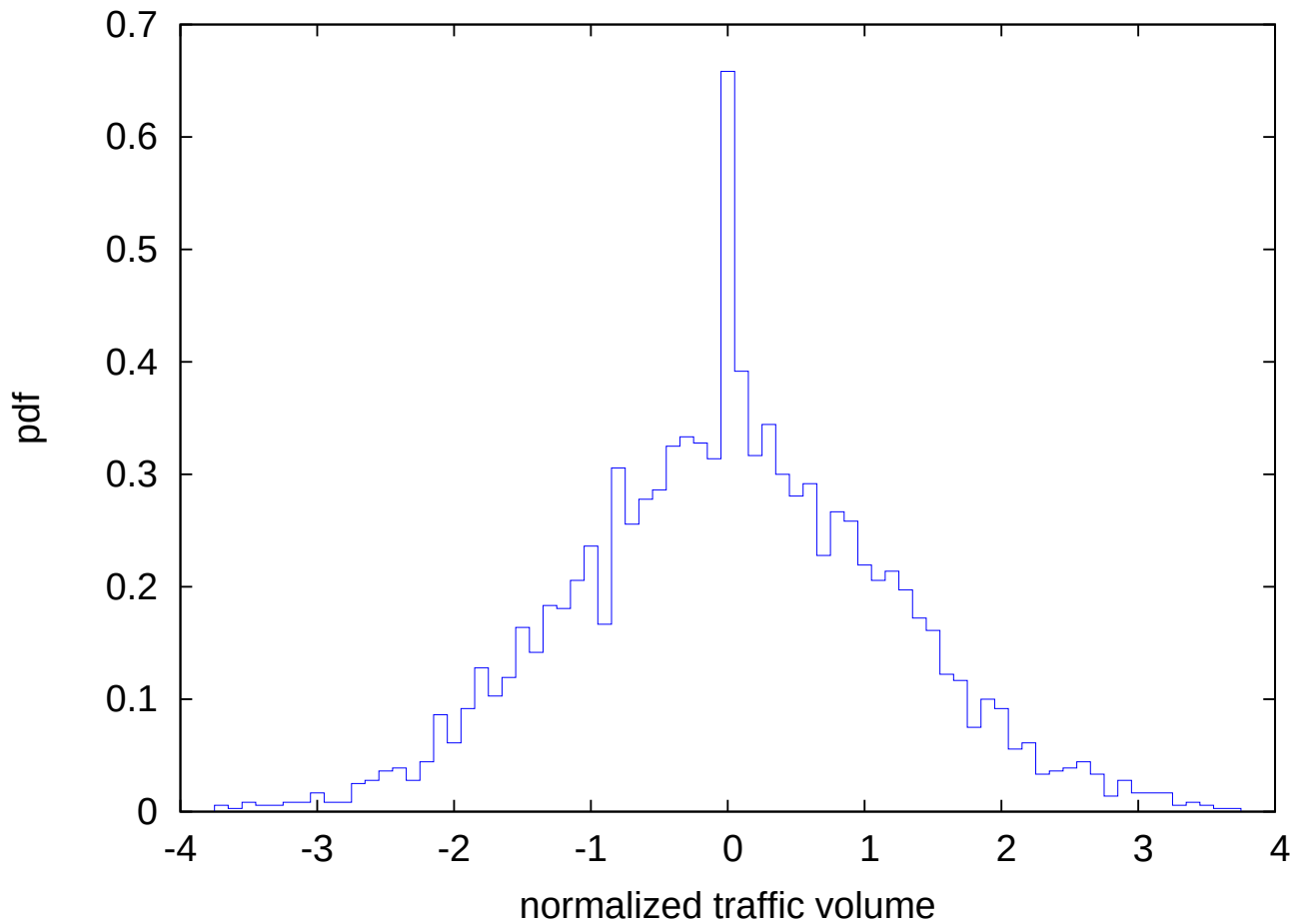
$$h = \frac{3.5\sigma}{n^{1/3}}$$

- あくまでも目安、分布の形状や変数の意味から読み易さを優先する

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 x 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)

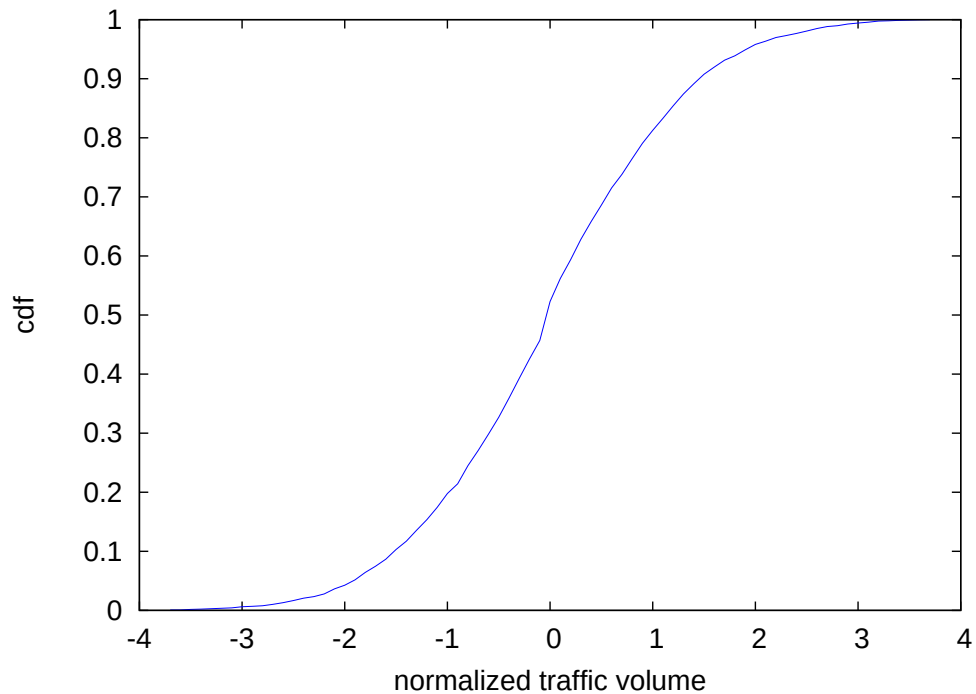
- 密度関数: x をいう値を観測する確率

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

- 累積分布関数: x 以下の値を観測する確率

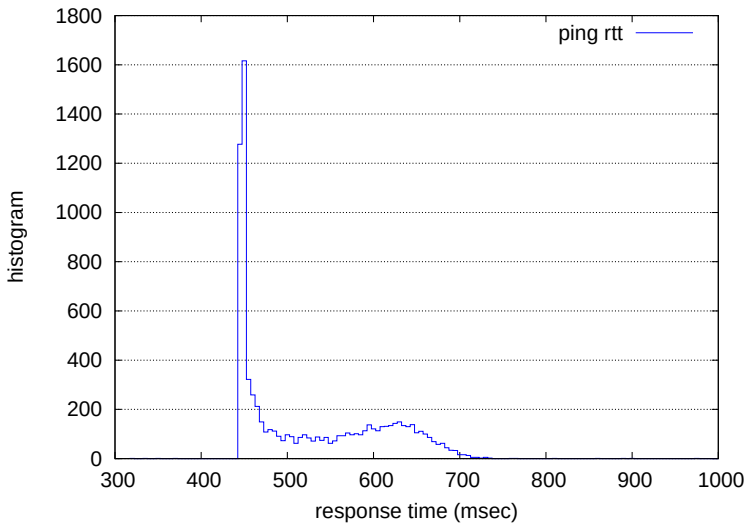
$$F(x) = P[X <= x]$$

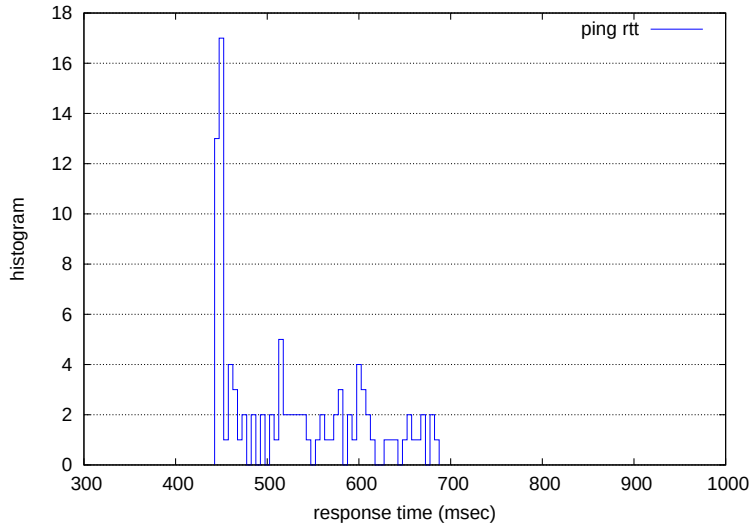
- 分布の偏りが大きい、サンプル数が少ない、外れ値が無視できない場合などは、ヒストグラムより有効

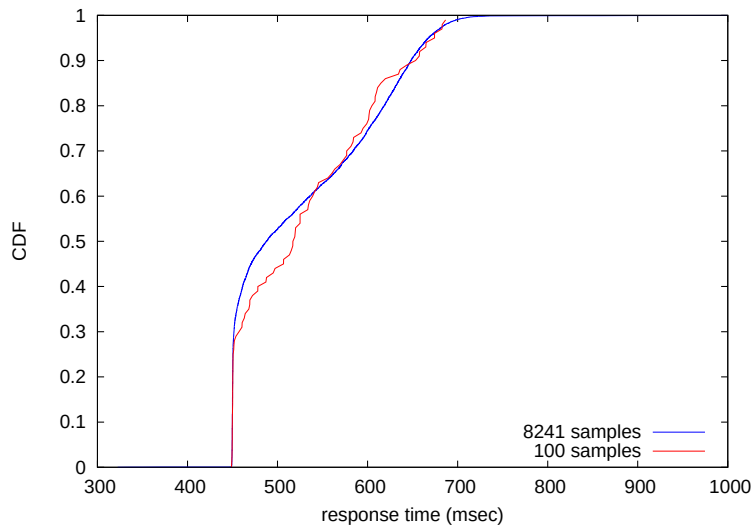


ヒストグラムと CDF の比較

- 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

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

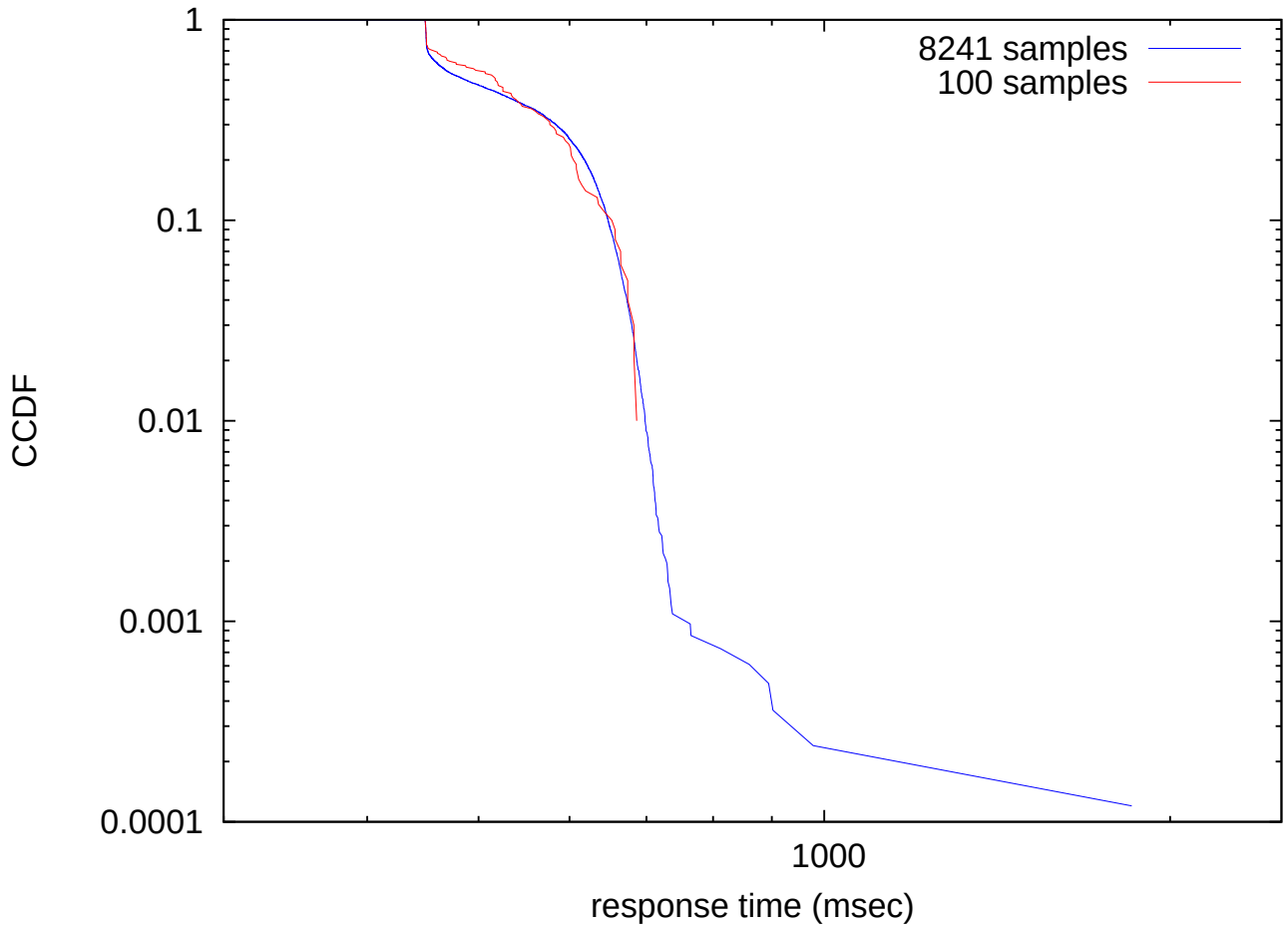
CCDF: x より大きい値の合計が全体に占める割合

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

- 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

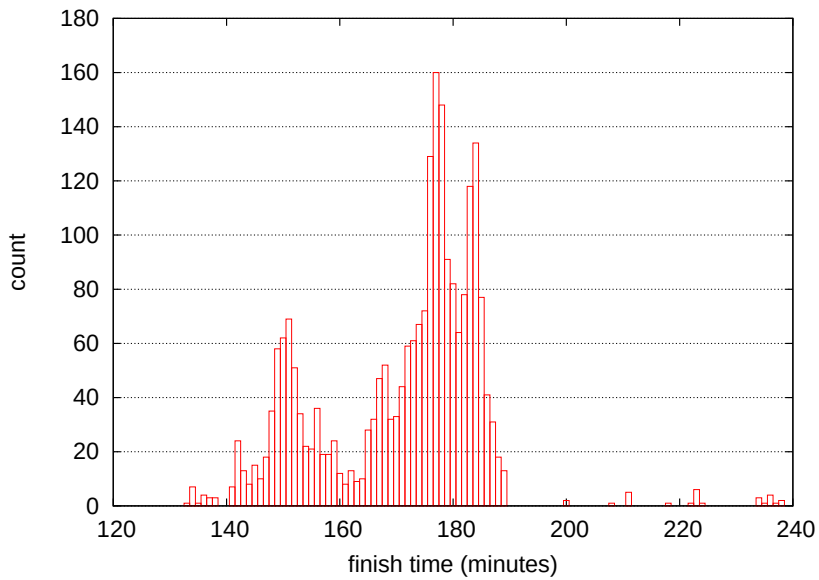
CCDF の例

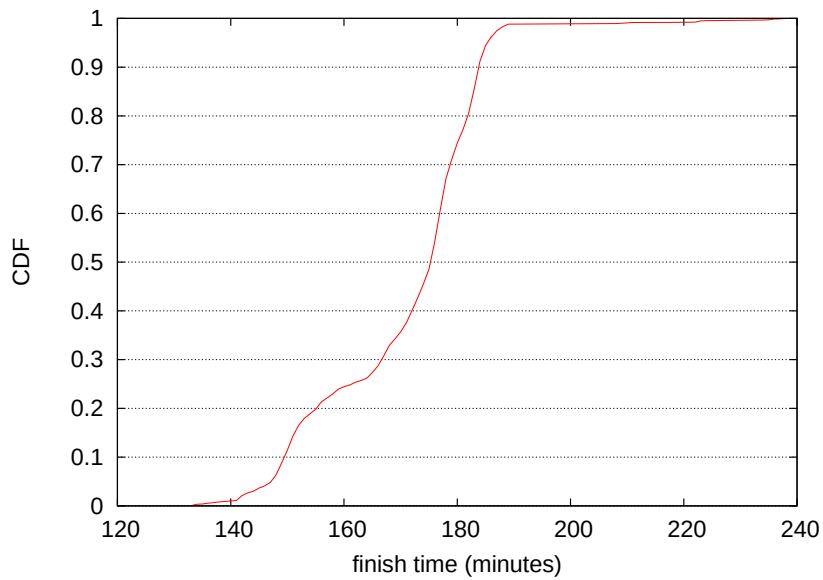
- 元データとサンプルのテイル部分の比較



演習: ヒストグラムと CDF

- ある市民マラソンの完走タイムの分布 (第 2 回授業演習)
- 今回は CDF をプロットする





演習: ヒストグラムと CDF (つづき)

- ある市民マラソンの完走タイムの分布 (第 2 回授業演習)
- 今回は CDF をプロットする

original:

```
# Minutes Count
133 1
134 7
135 1
136 4
137 3
138 3
141 7
142 24
...
```

add cumulative count:

```
# Minutes Count CumulativeCount
133 1 1
134 7 8
135 1 9
136 4 13
137 3 16
138 3 19
141 7 26
142 24 50
```

課題 2

- 課題 2: 最短経路木の計算、距離の分布と次数分布のプロット
- ねらい: トポロジーデータから特徴量を抽出しプロット
- データ: AS トポロジー (as-topology.txt)
  - CAIDA の skitter データから AS の隣接情報を抽出したデータ
- 提出項目
  1. 慶應 (38635) から他の AS への距離 (ホップ数) の分布のプロット
     - 慶應からの最短経路木を計算し、全ノードへの距離を求める
  2. この距離の分布を求めるスクリプト
     - 演習 2 のスクリプト実行結果を変換してもよい
  3. AS の次数分布の散布図
     - X 軸は次数 (ログスケール)、Y 軸はノード数
  4. AS の次数分布の CCDF プロット
     - X 軸は次数 (ログスケール)、Y 軸は CCDF
  5. この次数分布の CCDF を求めるスクリプト
- 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: 2011 年 7 月 8 日

課題 2 データ: AS トポロジー (as-topology.txt)

- AS 間の隣接関係 (23,660 nodes, 65,679 edges)
- エッジは無向、コストはすべて 1 としている

```
3 - 27724 1
3 - 291 1
3 - 3356 1
12 - 3754 1
13 - 668 1
14 - 3356 1
17 - 6503 1
18 - 6922 1
20 - 3356 1
22 - 668 1
24 - 127 1
25 - 2152 1
29 - 10578 1
32 - 174 1
...
```

まとめ

インターネットの多様性と複雑さを計る

- サンプリング
- 統計解析 (ヒストグラム、大数の法則)
- 演習:ヒストグラム、CDF
- **課題 2**

次回予定

第 8 回 ロングテールとさまざまな分布 (6/26)

- 正規分布
- その他の主要な分布
- 信頼区間と検定
- 演習:分布の生成、信頼区間