インターネット計測とデータ解析 第4回

長 健二朗

2011年6月1日

前回のおさらい

データの記録とログ解析

- 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

今日のテーマ

インターネットの速度を計る

- 速度計測
- 利用可能帯域の推測
- 平均 標準偏差
- 線形回帰
- 演習:平均、標準偏差、線形回帰
- **課題 1**

はじめに

速度とは？

- オブジェクトの単位時間あたりの位置変化

$$v = \frac{\Delta P}{\Delta t}$$

ネットワークの速度

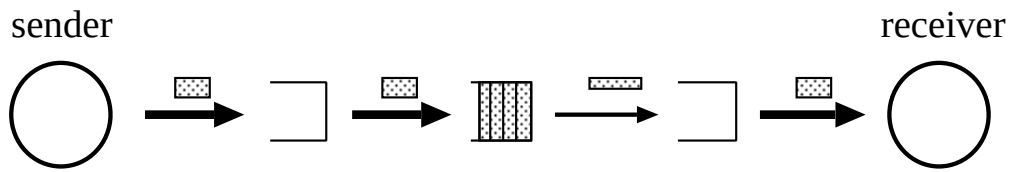
- 真空中の光の速度: $3.0x10^8 km/s$
- 光ファイバー中の伝播速度: $2.1x10^8 m/s$

データ転送レート: 正確には速度というより効率を表す

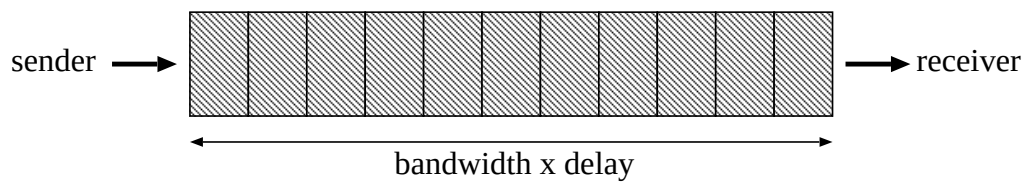
- ビットレート (bit rate): bits per second
- 関連用語 (とその混同)
  - 伝送容量 (bandwidth capacity)、帯域幅 (bandwidth)
  - スループット (throughput)

スループット計測

- 速度計測サイト: 実効転送速度を計測
  - ボトルネックはアクセス回線だと想定
    - 実際には ISP 境界などもボトルネックになる可能性
  - クロストラフィックの影響
- スループット計測ツール
  - Iperf、netperf、ixChariot など
- 輻輳: トラフィックの集中で品質低下する状態
  - 中間ルータのバッファにパケットが溜り遅延が増加

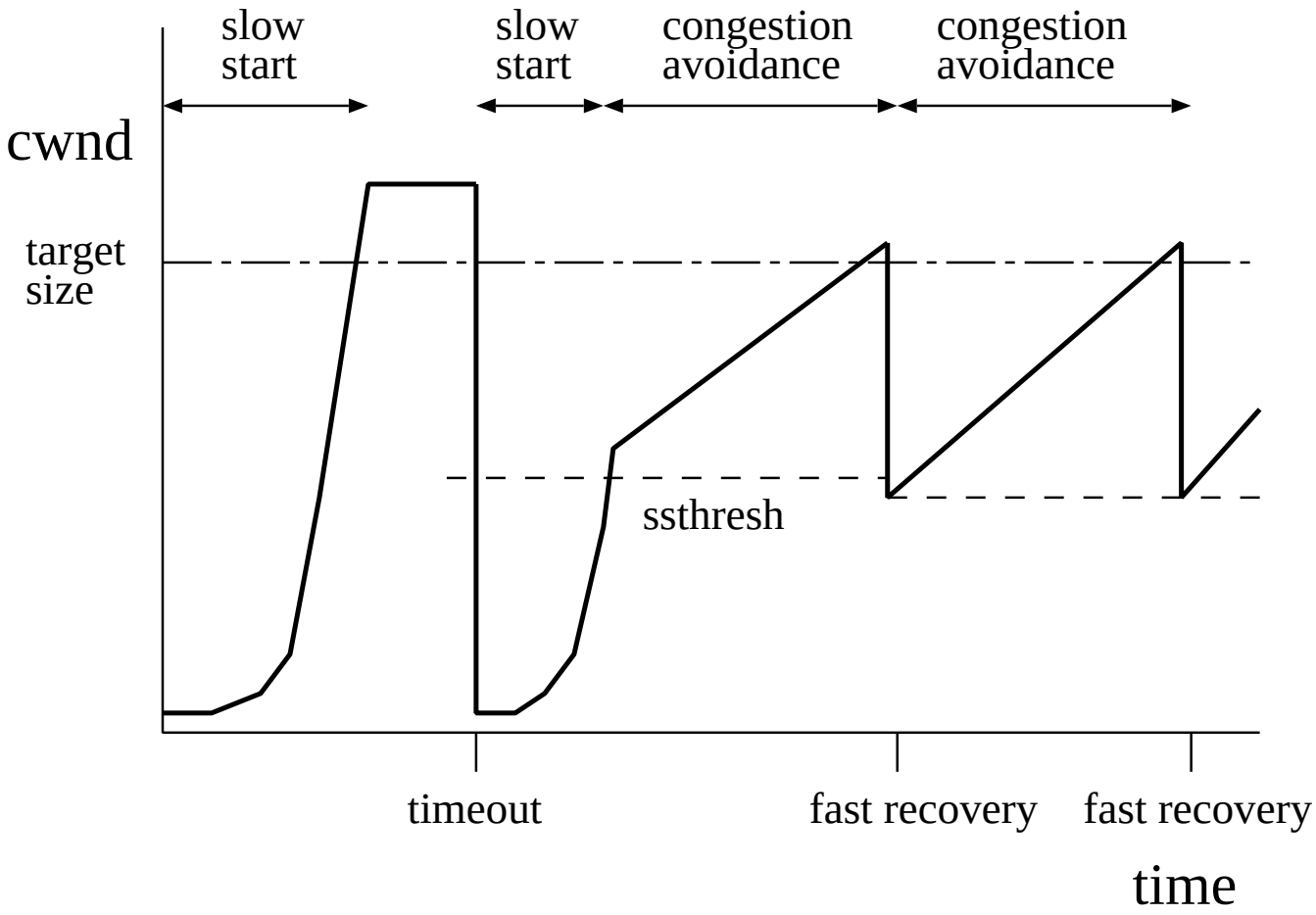


- TCP によるスループット計測
  - UDP ではバッファがオーバーフロー
  - TCP は利用可能な帯域に適応する
  - TCP バッファサイズ: delay bandwidth product



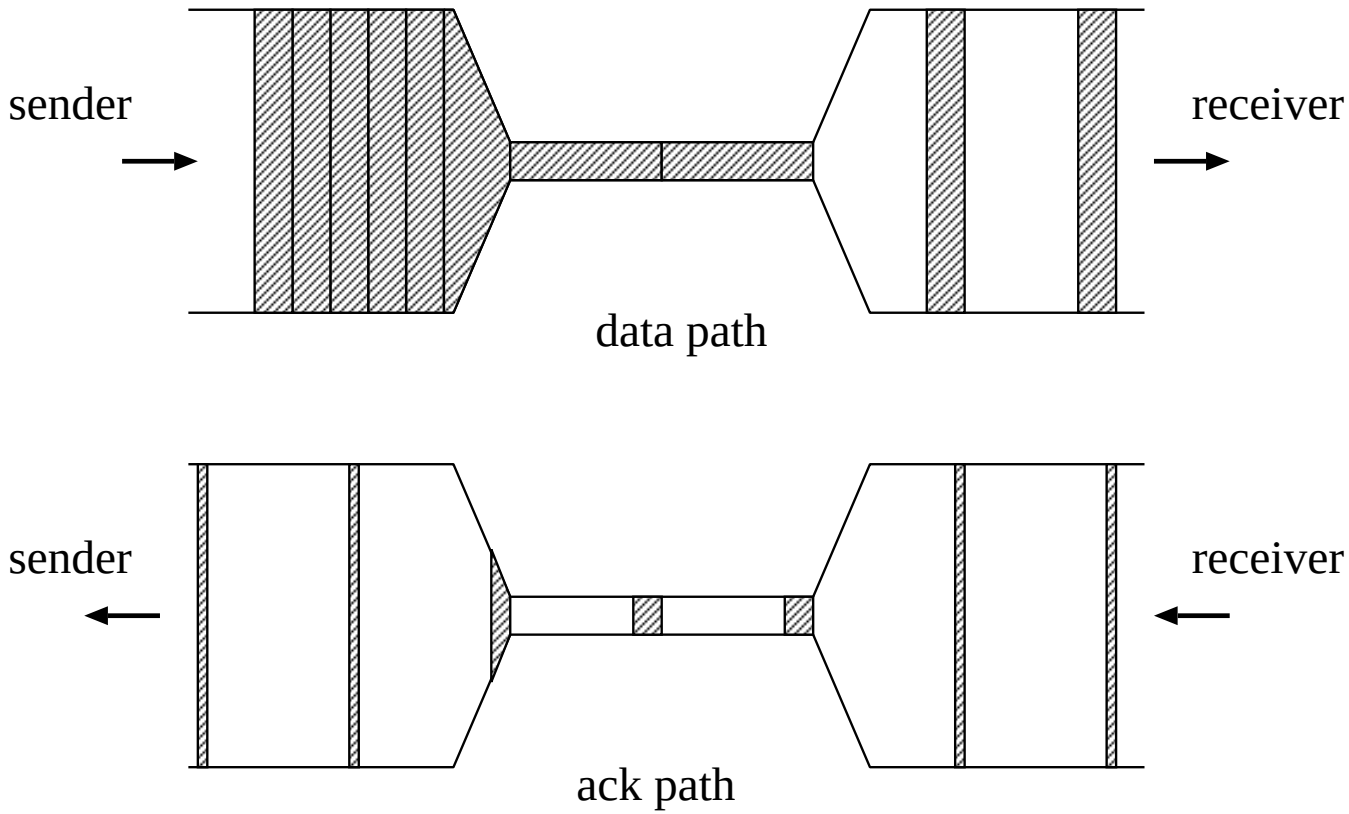
TCP congestion control

- congestion window により転送中のパケット量を調節
  - slow start/congestion avoidance
  - retransmit timeout
  - fast retransmit/fast recovery



TCP self-clocking

- ack の到着をトリガーに次のパケットを転送
- ボトルネック帯域に適応するメカニズム

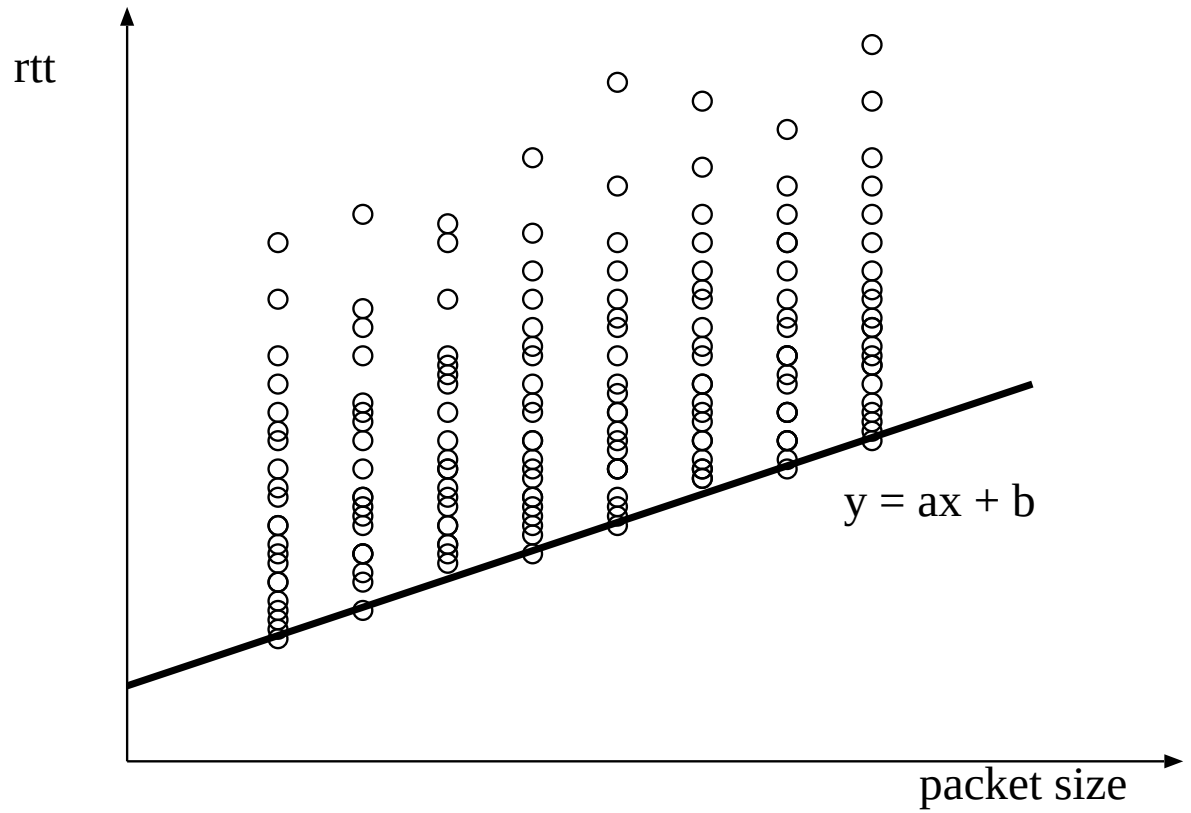


利用可能帯域の推測

- 回線を埋めることなく、利用可能帯域を推測する技術
- pathchar アルゴリズム (同様のツールが多数存在)
  - traceroute と同様に TTL を利用したホップ毎の遅延計測
  - 異なるパケットサイズで繰り返し計測
  - 各パケットサイズで最小遅延値を見つける
  - 線形回帰により伝送遅延と回線容量を推測
  - 手前のホップとの差分から、その区間の遅延と容量を求める
- 制約
  - 繰り返し計測の必要性
  - 誤差の累積: 特にボトルネック回線の先で精度低下
  - 一方向しか測定できない

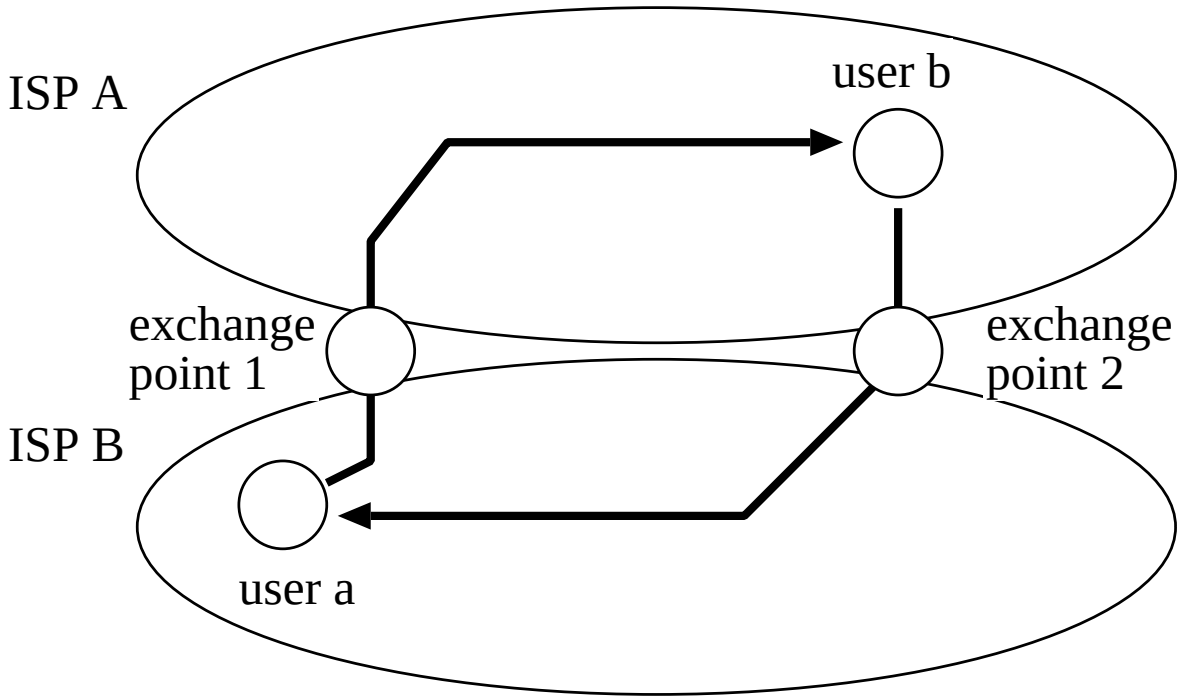
pathchar アルゴリズム

- 線形回帰による伝送遅延と回線容量の推測
  - $y = ax + b$
  - a: RTT-delta/packetsize-delta (bandwidth)
  - b: delay for packetsize 0 (propagation delay)



非対称な経路

- 事業者間の非対称経路は一般的
  - 通常、最も近い接続点で相手事業者にパケットを転送



要約統計量 (summary statistics)

標本の分布の特徴を要約して表す数値

- 位置を表す数値:
  - 平均 (mean)、中央値 (median)、最頻値 (mode)
- ばらつきを表す数値:
  - 範囲 (range)、分散 (variance)、標準偏差 (standard deviation)

位置を表す数値

- 平均 (mean):

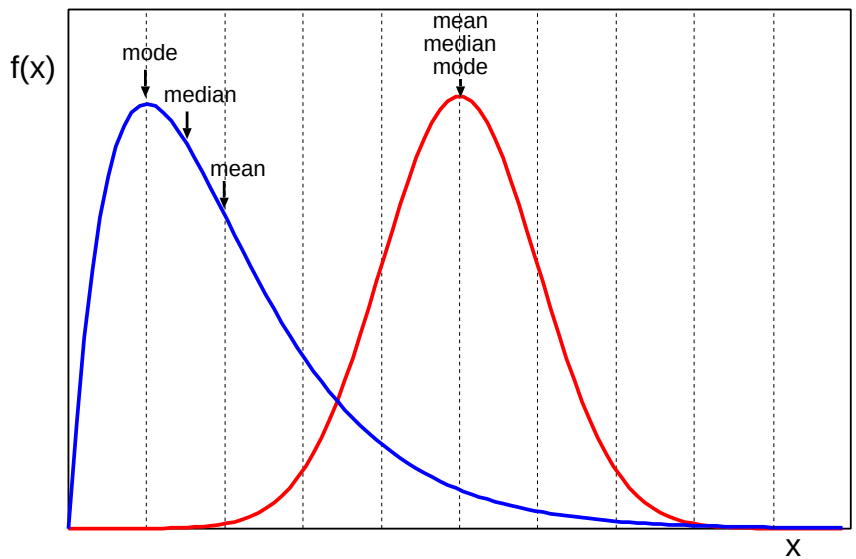
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 中央値 (median): データの値をソートして中央にくる値

$$x_{median} = \begin{cases} x_{r+1} & m \text{ が奇数の場合, } m = 2r+1 \\ (x_r + x_{r+1})/2 & m \text{ が偶数の場合, } m = 2r \end{cases}$$

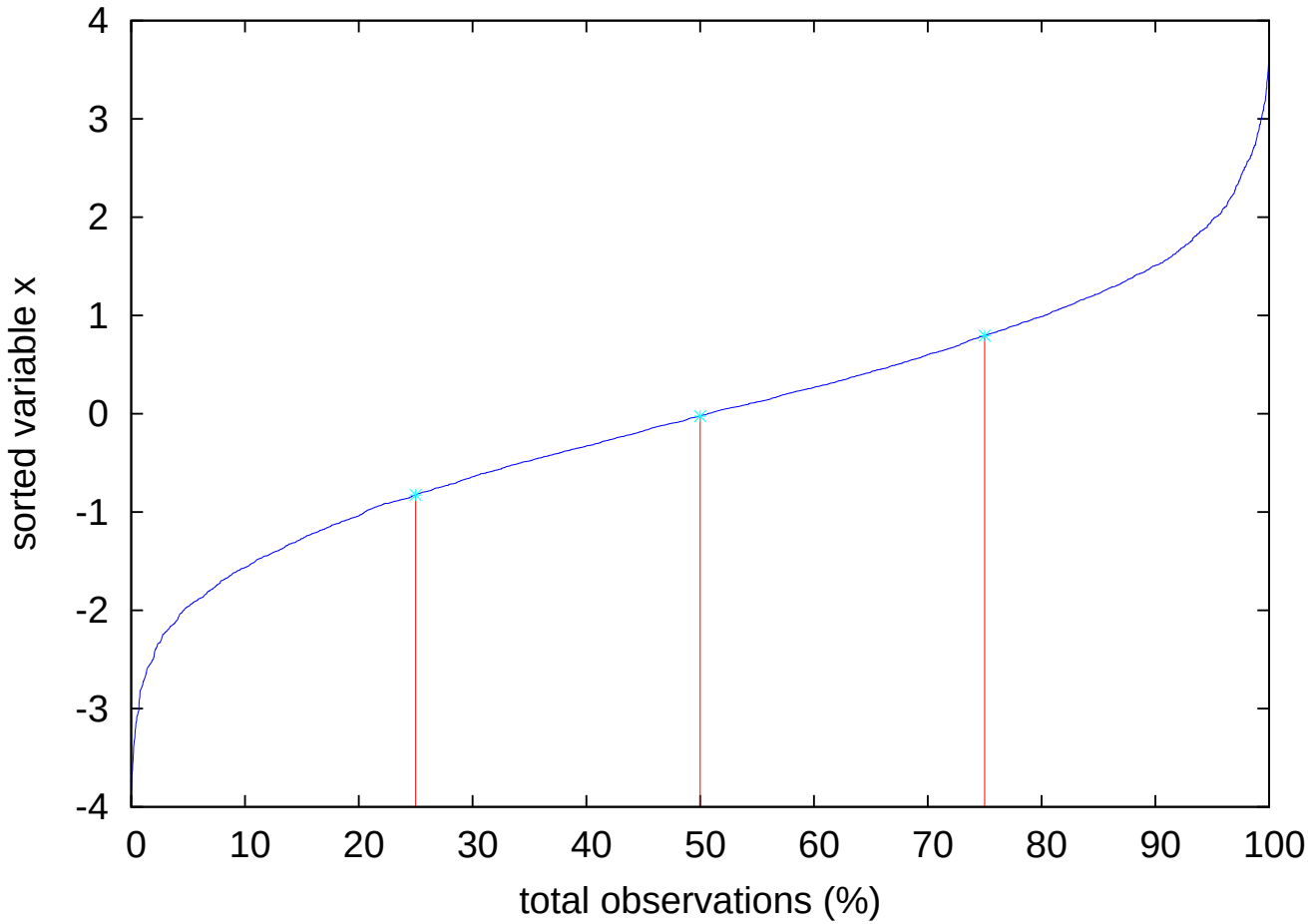
- 最頻値 (mode): 出現頻度が最も高い値

対称な分布であれば、これらは同一



パーセンタイル (percentiles)

- pth-percentile: 小さい方から数えて p%目の値
  - median = 50th-percentile

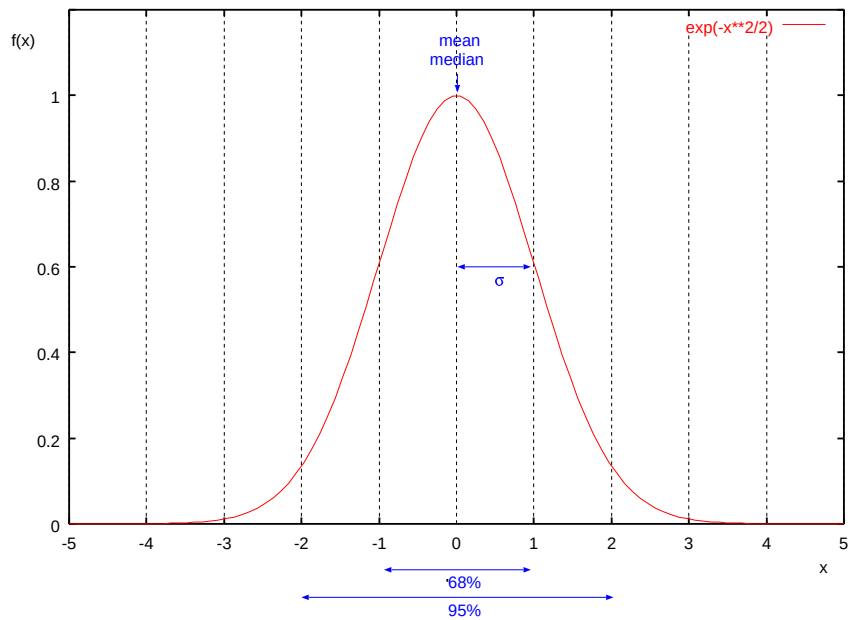


ばらつきを表す数値

- 範囲 (range): 最大値と最小値の差
- 分散 (variance):

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

- 標準偏差 (standatd deviation): σ
  - 平均と同じ次元なので直接比較可能
  - 統計的なばらつきを示すのに最も良く使われる値
- 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



相関 (correlation)

- 共分散 (covariance):

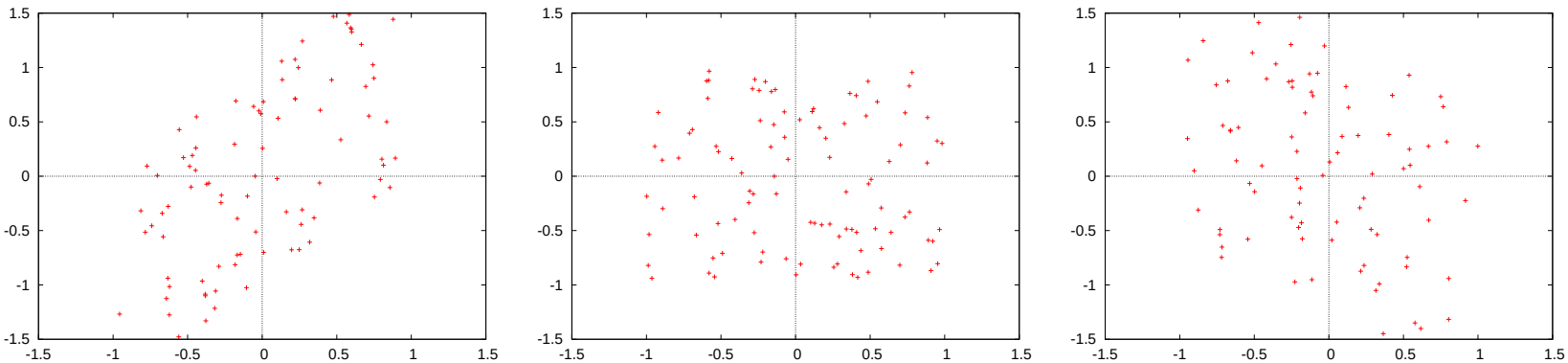
$$\sigma_{xy}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- 相関係数 (correlation coefficient):

$$\rho_{xy} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

散布図 (scatter plots)

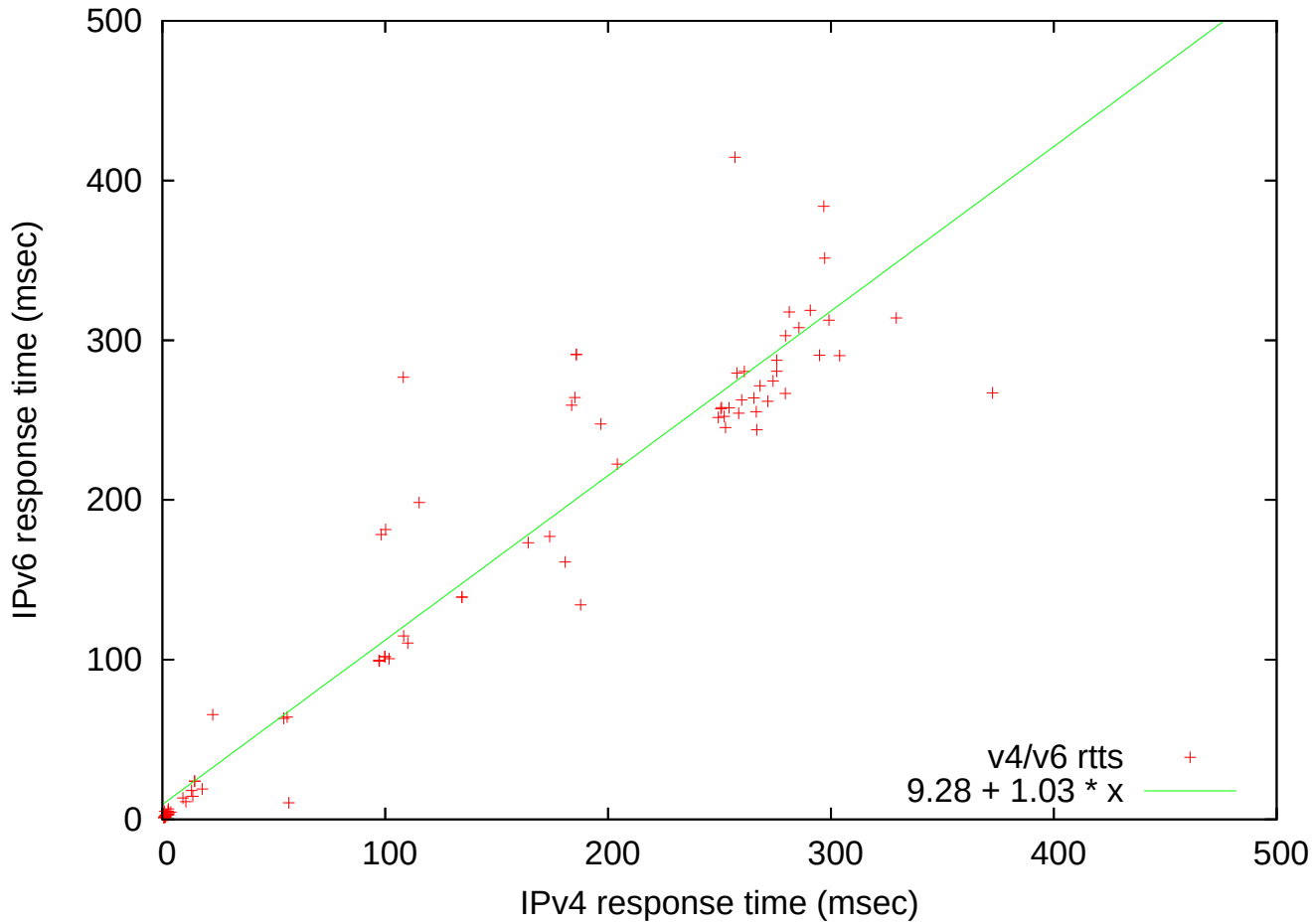
- 2 つの変数の関係を見るのに有効
  - X 軸: 変数 X
  - Y 軸: それに対応する変数 Y の値
- 散布図で分かる事
  - X と Y に関連があるか
    - 無相関、正の相関、負の相関
  - 変数 Y が変数 X に依存して変化するかどうか
  - 外れ値の存在があるか



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

線形回帰 (linear regression)

- データに一次関数を当てはめる
  - 最小二乗法 (least square method): 誤差の二乗和を最小にする



最小二乗法 (least square method)

誤差の二乗和を最小にする一次関数を求める

$$f(x) = b_0 + b_1 x$$

切片と傾きの求め方

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

ここで

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

$$\sum xy = \sum_{i=1}^{n} x_i y_i \qquad \sum x^2 = \sum_{i=1}^{n} (x_i)^2$$

最小二乗法の導出

i 番目の変数の誤差 $e_i = y_i - (b_0 + b_1 x_i)$、$n$ 回の観測における誤差の平均は

$$\bar{e} = \frac{1}{n}\sum_i e_i = \frac{1}{n}\sum_i (y_i - (b_0 + b_1 x_i)) = \bar{y} - b_0 - b_1 \bar{x}$$

誤差平均が 0 になるようにすると $b_0 = \bar{y} - b_1 \bar{x}$
b_0 を b_1 で表現すると $e_i = y_i - \bar{y} + b_1 \bar{x} - b_1 x_i = (y_i - \bar{y}) - b_1 (x_i - \bar{x})$
誤差の二乗和 SSE は

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [(y_i - \bar{y})^2 - 2b_1 (y_i - \bar{y})(x_i - \bar{x}) + b_1^2 (x_i - \bar{x})^2]$$

分散に書き直す

$$\begin{aligned} \frac{SSE}{n} &= \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2 - 2b_1 \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) + b_1^2 \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= \sigma_y^2 - 2b_1 \sigma_{xy}^2 + b_1^2 \sigma_x^2 \end{aligned}$$

SSE を最小にする b_1 は、この式を b_1 の 2 次式とみて b_1 について微分して 0 と置く

$$\frac{1}{n}\frac{d(SSE)}{db_1} = -2\sigma_{xy}^2 + 2b_1 \sigma_x^2 = 0$$

すなわち $b_1 = \frac{\sigma_{xy}^2}{\sigma_x^2} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2}$

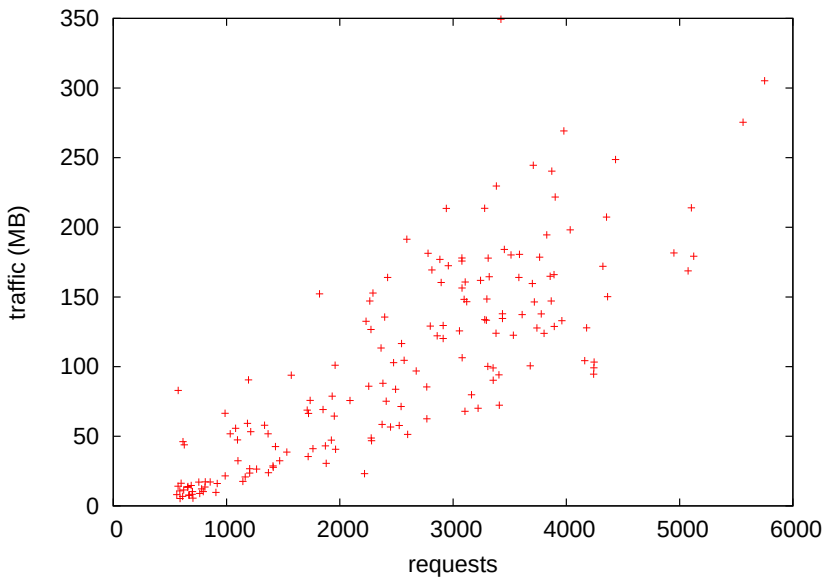
演習:平均、標準偏差、線形回帰

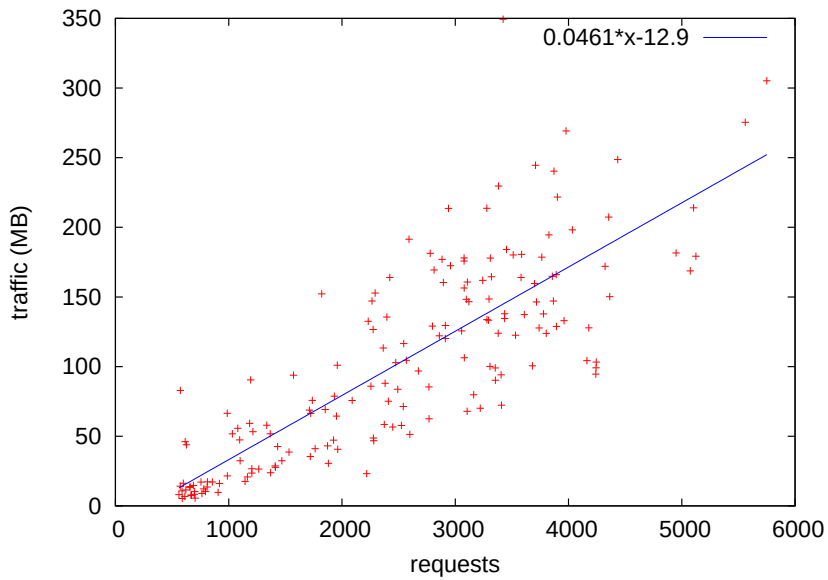
平均と標準偏差の計算

```
# extract a variable from each line
re = /^\S+\s+(\d+)\s+\d+/

# create an array for data
data = Array.new
ARGF.each_line do |line|
  if re.match(line)
    data.push $1.to_i
  end
end

# compute mean
sum = 0
data.each {|v| sum += v}
mean = Float(sum) / data.length

# compute standard deviation
sqsum = 0
data.each {|v| sqsum += (v - mean)**2}
var = Float(sqsum) /data.length
stddev = Math.sqrt(var)

puts "mean=#{mean} stddev=#{stddev}"
```

線形回帰

- １時間ごとのリクエスト数と転送数の関係
- 最小二乗法で回帰直線を求める





線形回帰コード

```
# extract 2 variables from each line
re = /^\S+\s+(\d+)\s+(\d+)/

# compute (y = b0 + b1*x) by the least square method
sum_x = sum_y = sum_xx = sum_xy = n = 0
ARGF.each_line do |line|
    if re.match(line)
      x = $1.to_i
      y = $2.to_i

      sum_x += x
      sum_y += y
      sum_xx += x * x
      sum_xy += x * y
      n += 1
    end
end

mean_x = Float(sum_x) / n
mean_y = Float(sum_y) / n
b1 = (sum_xy - n * mean_x * mean_y) / (sum_xx - n * mean_x * mean_x)
b0 = mean_y - b1 * mean_x

puts "b0=#{b0} b1=#{b1}"
```

課題 1

- 課題: 平均・標準偏差の計算と結果のプロット
- ねらい: 統計処理のプログラミング、プロット作成
- データ: 2011-05-16/2011-05-22 のアクセスログ (SFC-FSF から入手する)
  - 2011-02-28/2011-03-06 は演習用、2011-05-16/2011-05-22 は課題用
  - 注意: 説明には、2011-02-28/2011-03-06 のデータを使っているため、結果は少し異なる
- 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
  1. プロット用データ作成スクリプト
  2. プロット用データテーブル
  3. 各曜日の時間別リクエスト数プロット
  4. 全体の時間別リクエスト数の平均と標準偏差プロット
- 提出〆切: 2011 年 6 月 17 日

プロット用データテーブルの作成

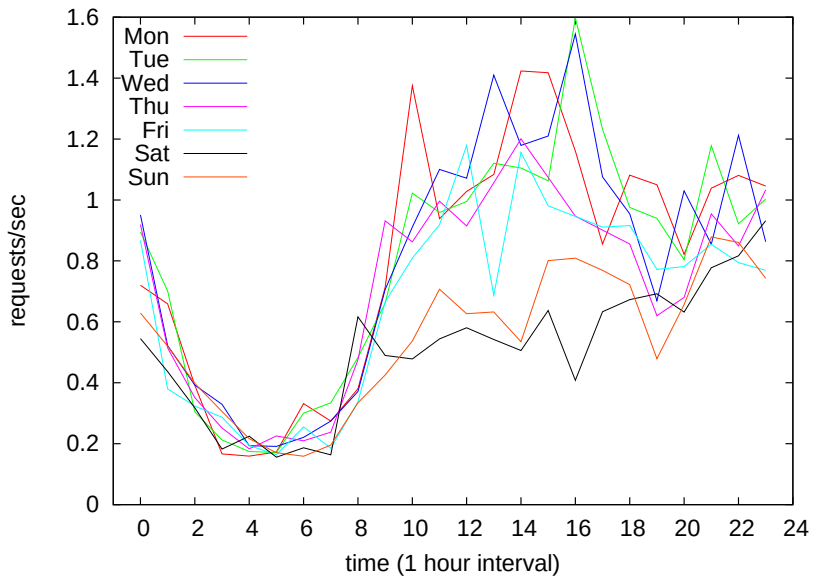
- 第 3 回の演習で作成した 1 時間ごとのリクエスト数と転送量を抽出するスクリプトを用いて、2011-05-16/2011-05-22 の 1 時間ごとのデータを作る
- このデータから、以下のようなプロット用データテーブルを作成する
  1. プロット用データテーブル作成スクリプトと出来たデータテーブルを提出

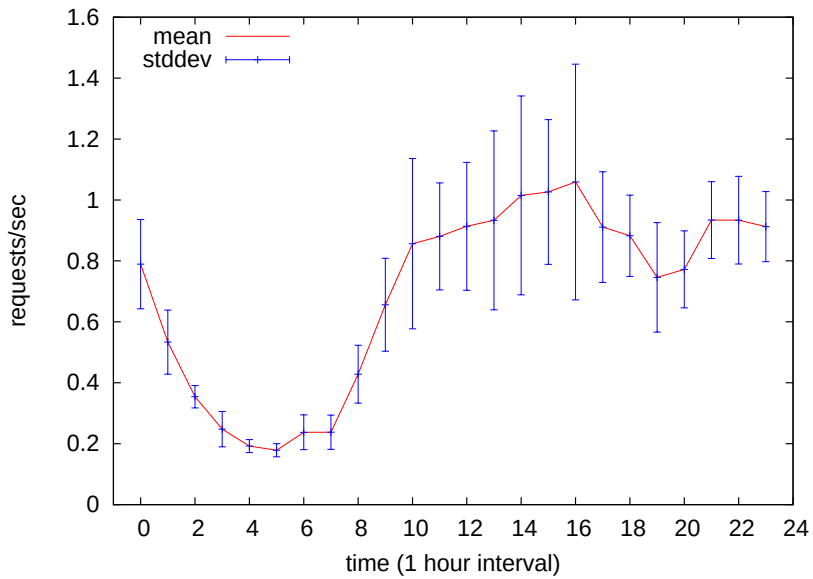
```
#hour Mon  Tue  Wed  Thu  Fri  Sat  Sun  mean   stddev
 0    2592 3221 3423 3310 3121 1963 2265 2842.1 527.27
 1    2373 2525 1880 1852 1367 1572 1873 1920.2 379.20
 2    1410 1097 1412 1265 1165 1144 1432 1275.0 132.38
 3     599  764 1185  906 1033  657 1101  892.1 209.44
...
22    3891 3319 4364 3056 2860 2940 3098 3361.1 518.11
23    3764 3610 3107 3718 2768 3354 2675 3285.1 413.76
```

グラフ

プロット用データテーブルから以下の 2 種類のプロットを作成

- 各曜日の時間別リクエスト数プロット
- 時間別に曜日別リクエスト数の平均をプロットし、標準偏差をエラーバーで示す





まとめ

インターネットの速度を計る

- 速度計測
- 利用可能帯域の推測
- 平均 標準偏差
- 線形回帰
- 演習:平均、標準偏差、線形回帰
- **課題 1**

次回予定

第 5 回 インターネットの構造を計る (6/8)

- インターネットアーキテクチャ
- ネットワーク階層
- トポロジー
- グラフ理論
- 演習:トポロジ解析