インターネット計測とデータ解析 第8回

長 健二朗

2011年6月26日

前回のおさらい

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**

今日のテーマ

ロングテールとさまざまな分布

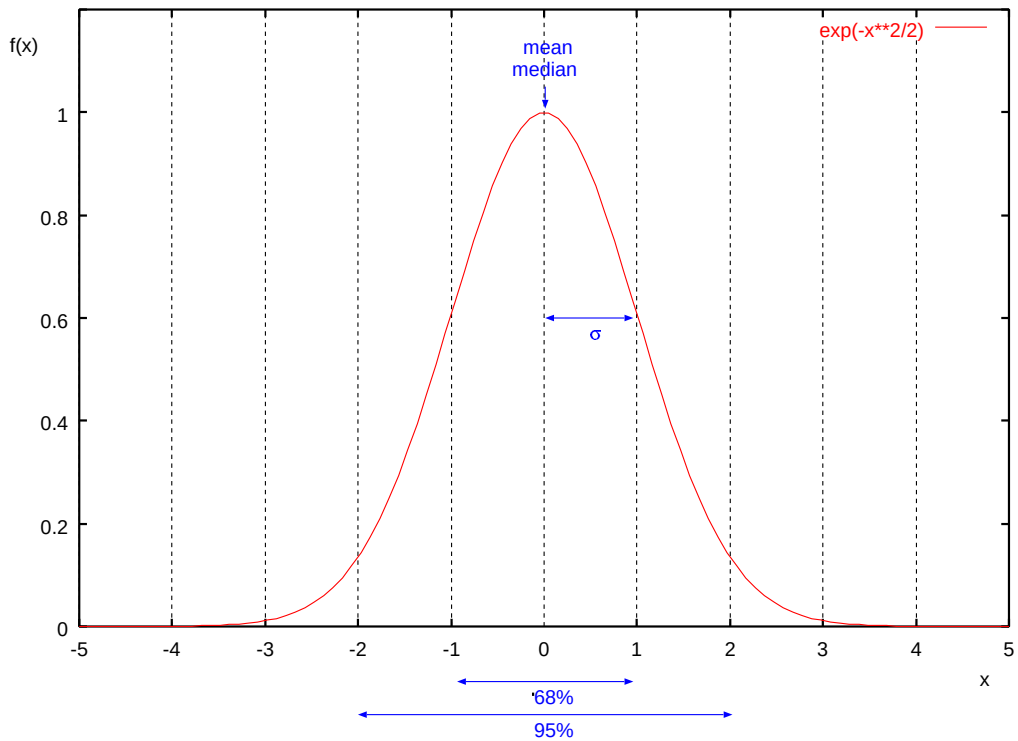
- 正規分布
- その他の主要な分布
- 信頼区間と検定
- 演習:分布の生成、信頼区間

さまざまな分布

- 正規分布
- 指数分布
- べき分布

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

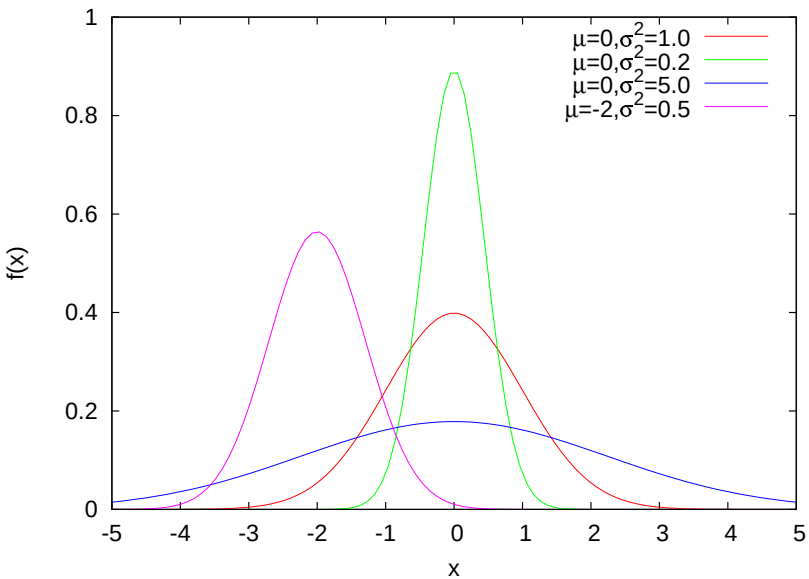
確率密度関数 (PDF)

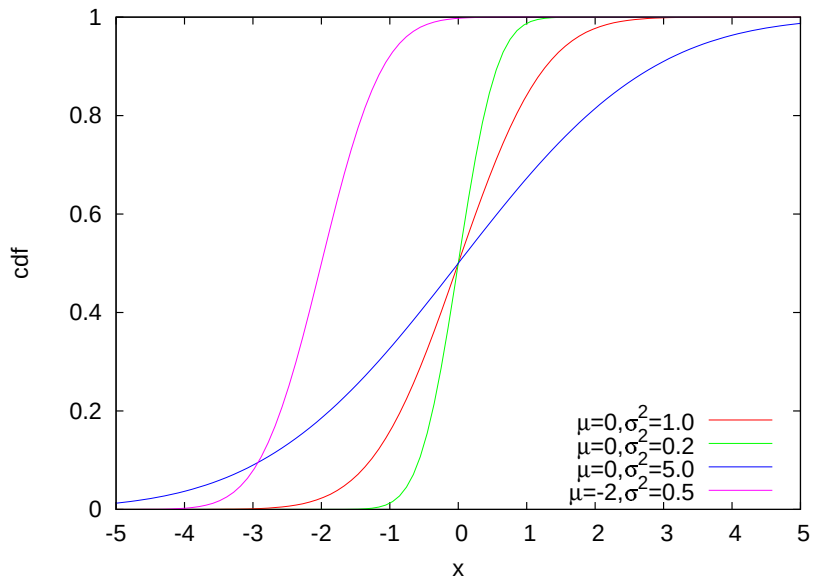
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

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)

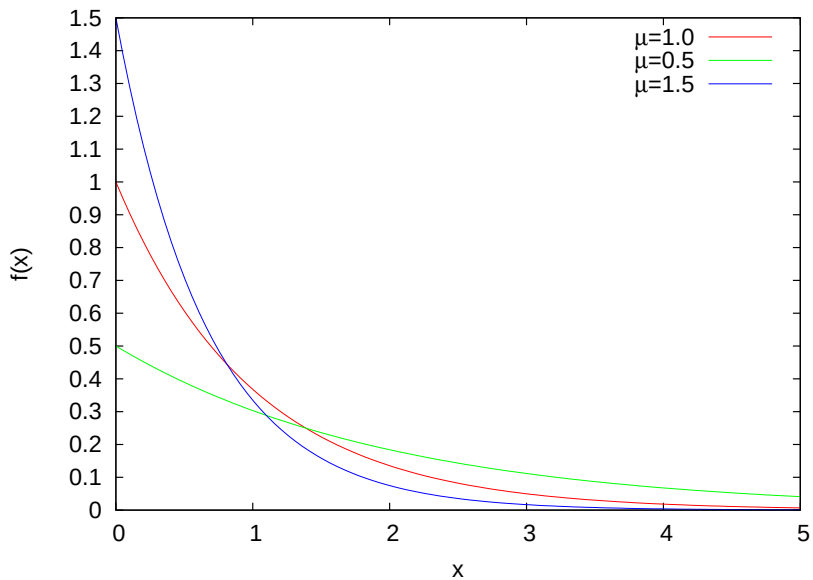
$$f(x) = \lambda e^{-\lambda x}, (x \geq 0)$$

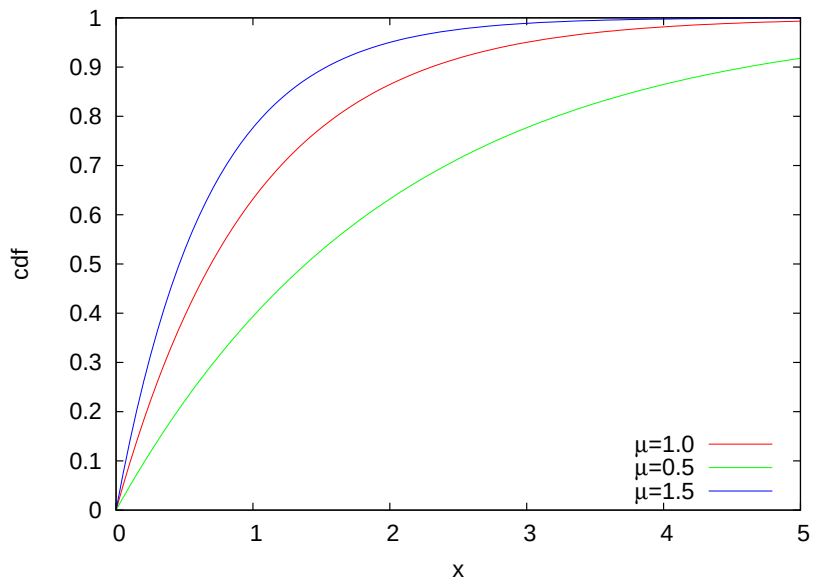
累積分布関数 (CDF)

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

$$\lambda > 0 : \textit{rate parameter}$$

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





べき分布 (power-law distribution)

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

パレート分布: ネットワーク研究で最も使われる

パレート分布 (pareto distribution)

確率密度関数 (PDF)

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

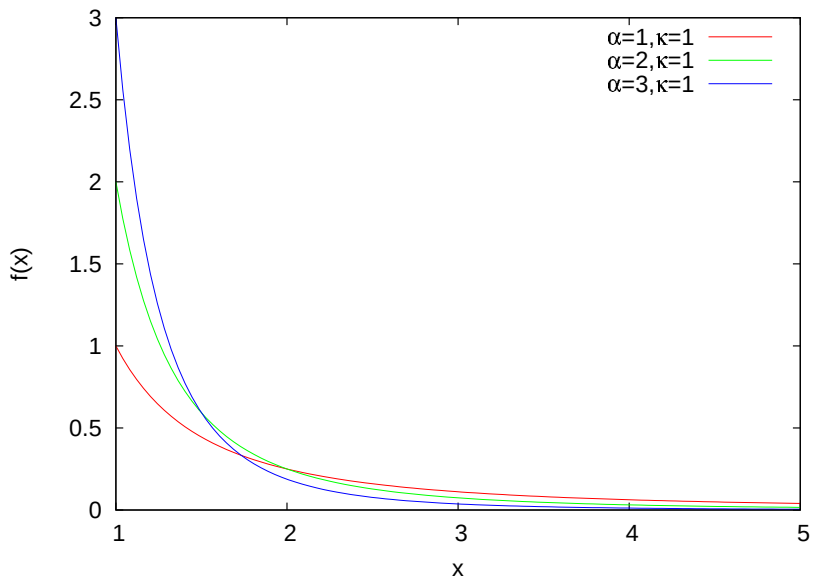
累積分布関数 (CDF)

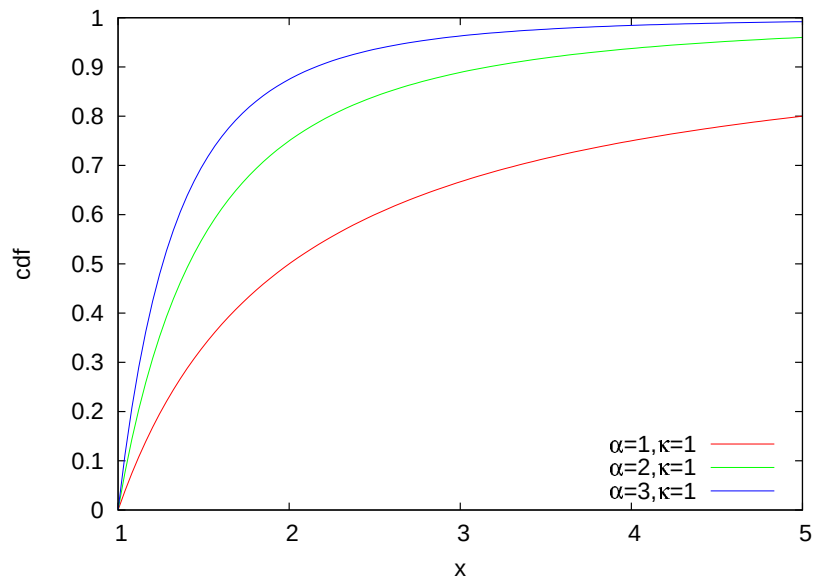
$$F(x) = 1 - (\frac{\kappa}{x})^{\alpha}$$

κ : *minimum value of* x, α : *pareto index*

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

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





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

CCDF のプロット

CDF のプロット

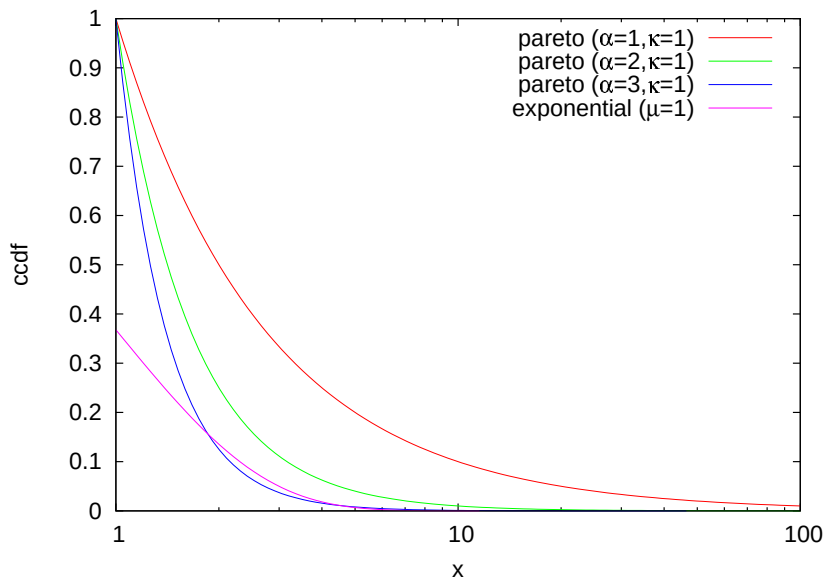
- $x_i, i \in \{1, \ldots, n\}$ を値順にソート
- $(x_i, \frac{1}{n}\sum_{k=1}^{i} k)$ をプロット
- Y 軸は通常リニアスケール

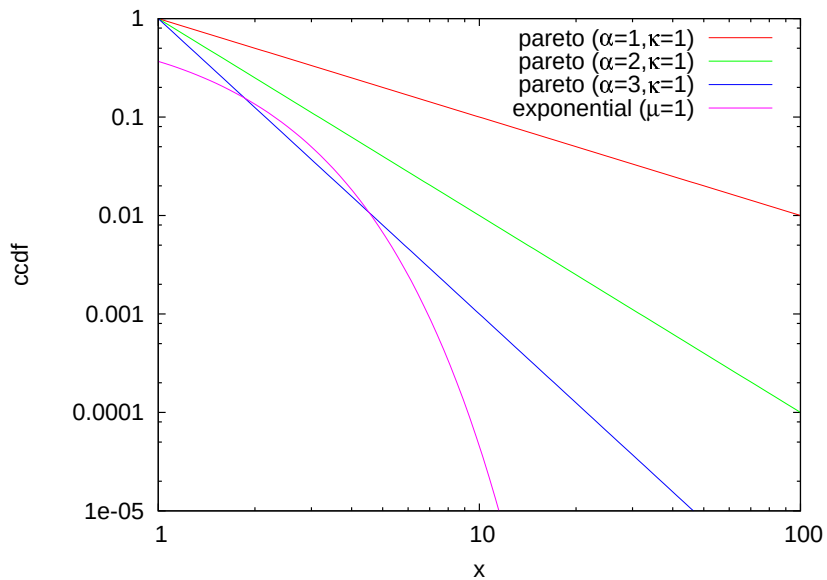
CCDF のプロット

- $x_i, i \in \{1, \ldots, n\}$ を値順にソート
- $(x_i, 1 - \frac{1}{n}\sum_{k=1}^{i} k)$ 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





信頼区間 (confidence interval)

- 信頼区間 (confidence interval)
  - 統計的に真値に範囲を示す
  - 推定値の確かさ、不確かさを示す
- 信頼度 (confidence level) 有意水準 (significance level)

$$Prob\{c_1 \leq \mu \leq c_2\} = 1 - \alpha$$

$(c1, c2)$: *confidence interval*
$100(1-\alpha)$: *confidence level*
α : *significance level*

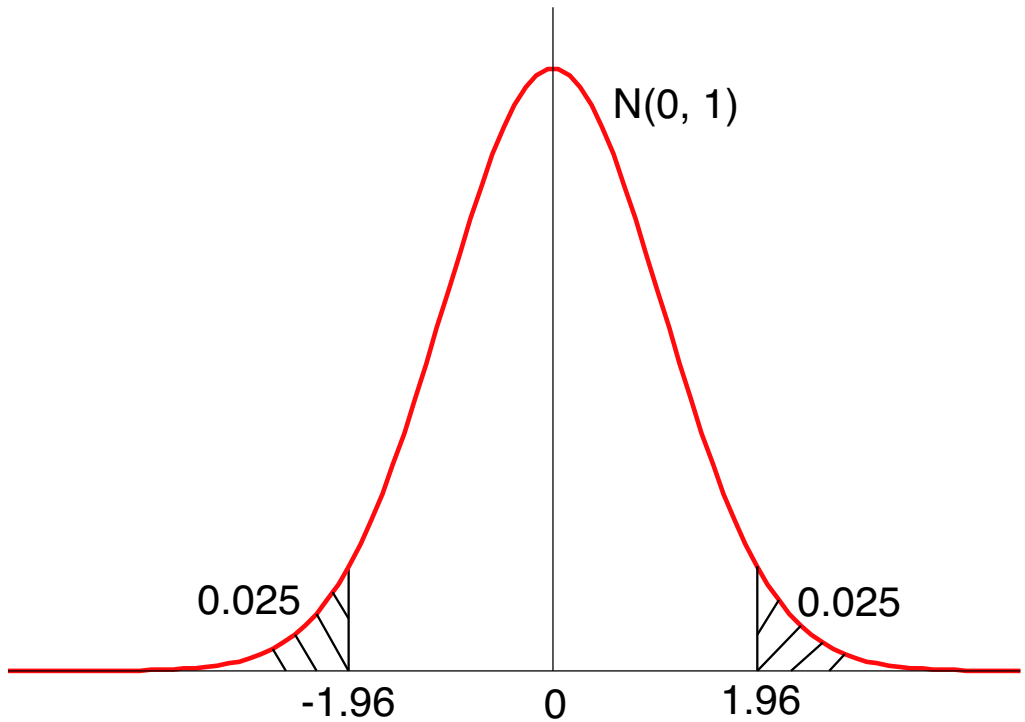
- 例: 信頼度 95% で、母平均は、$c1$ と $c2$ の間に存在
- 慣習として、信頼度 95% と 99% がよく使われる

95%信頼区間

正規母集団 $N(\mu, \sigma)$ から得られた標本平均 $\bar{x}$ は正規分布 $N(\mu, \sigma/\sqrt{n})$ に従う

95%信頼区間は標準正規分布の以下の部分を意味する

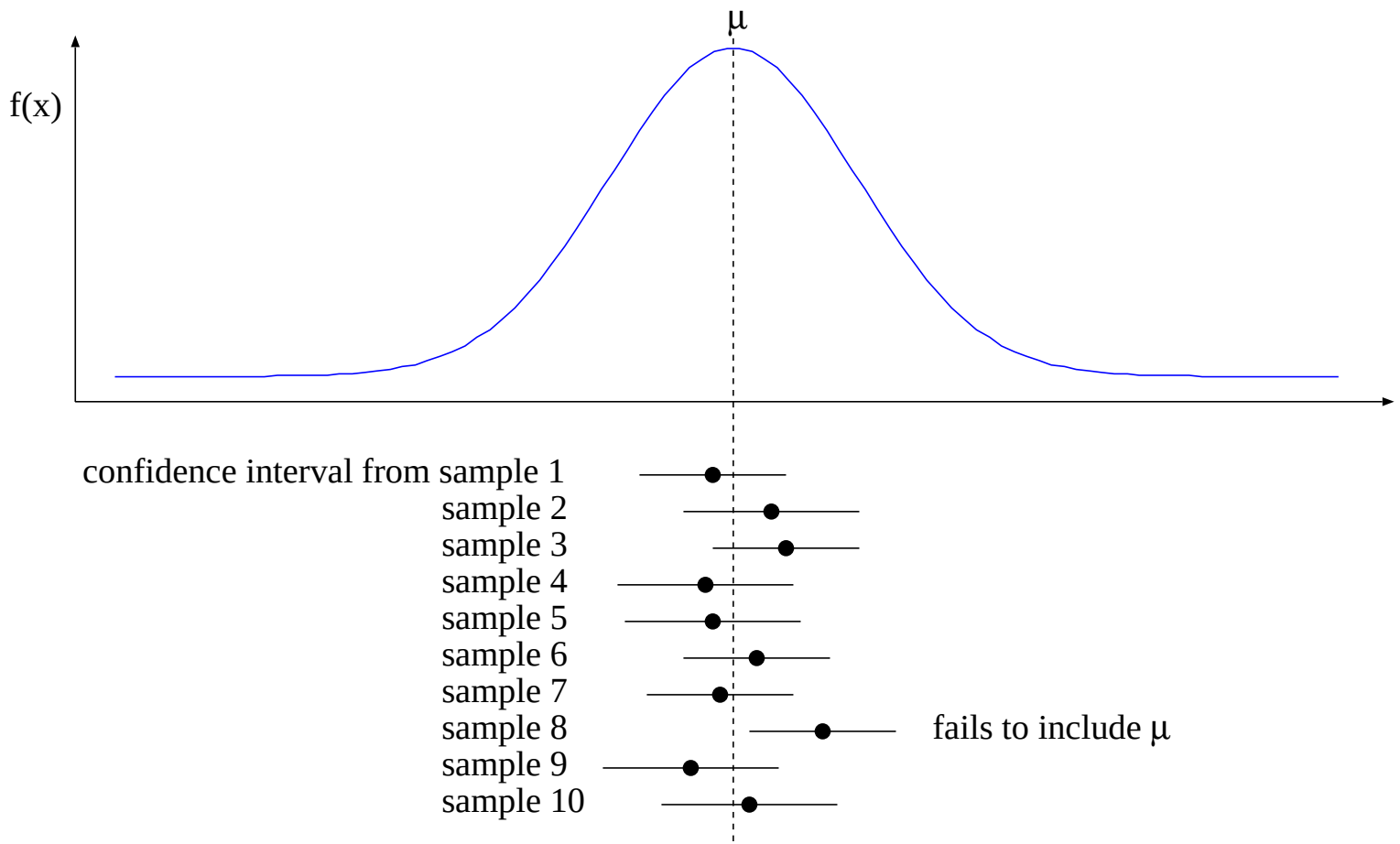
$$-1.96 \leq \frac{\bar{x} - \mu}{\sigma\sqrt{n}} \leq 1.96$$



標準正規分布 N(0, 1)

信頼区間の意味

- 信頼度 90% とは、90% の確率で母平均が信頼区間内に存在すること



平均値の信頼区間

サンプルサイズが大きければ、母平均の信頼区間は、

$$\bar{x} \mp z_{1-\alpha/2}\, s/\sqrt{n}$$

ここで、$\bar{x}$:標本平均 s:標本標準偏差 n:標本数 α:有意水準
$z_{1-\alpha/2}$:標準正規分布における $(1-\alpha/2)$ 領域の境界値

- 信頼度 95% の場合: $z_{1-0.05/2} = 1.960$
- 信頼度 90% の場合: $z_{1-0.10/2} = 1.645$
- 例: TCP スループットを 5 回計測
  - 3.2, 3.4, 3.6, 3.6, 4.0Mbps
  - 標本平均:$\bar{x} = 3.56$Mbps 標本標準偏差:$s = 0.30$Mbps
  - 95%信頼区間:

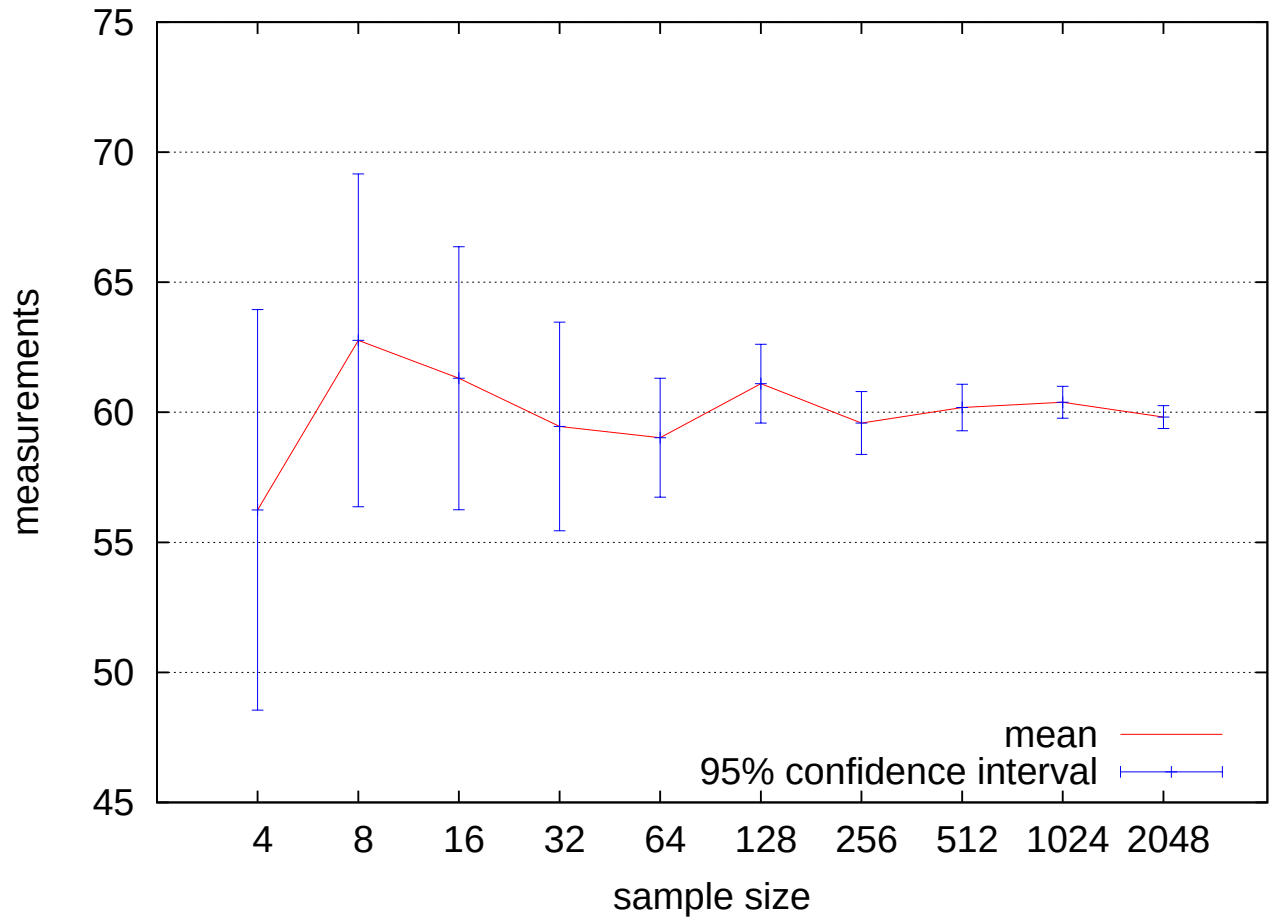
    $$\bar{x} \mp 1.96(s/\sqrt{n}) = 3.56 \mp 1.960 \times 0.30/\sqrt{5} = 3.56 \mp 0.26$$

  - 90%信頼区間:

    $$\bar{x} \mp 1.645(s/\sqrt{n}) = 3.56 \mp 1.645 \times 0.30/\sqrt{5} = 3.56 \mp 0.22$$

平均値の信頼区間とサンプル数

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

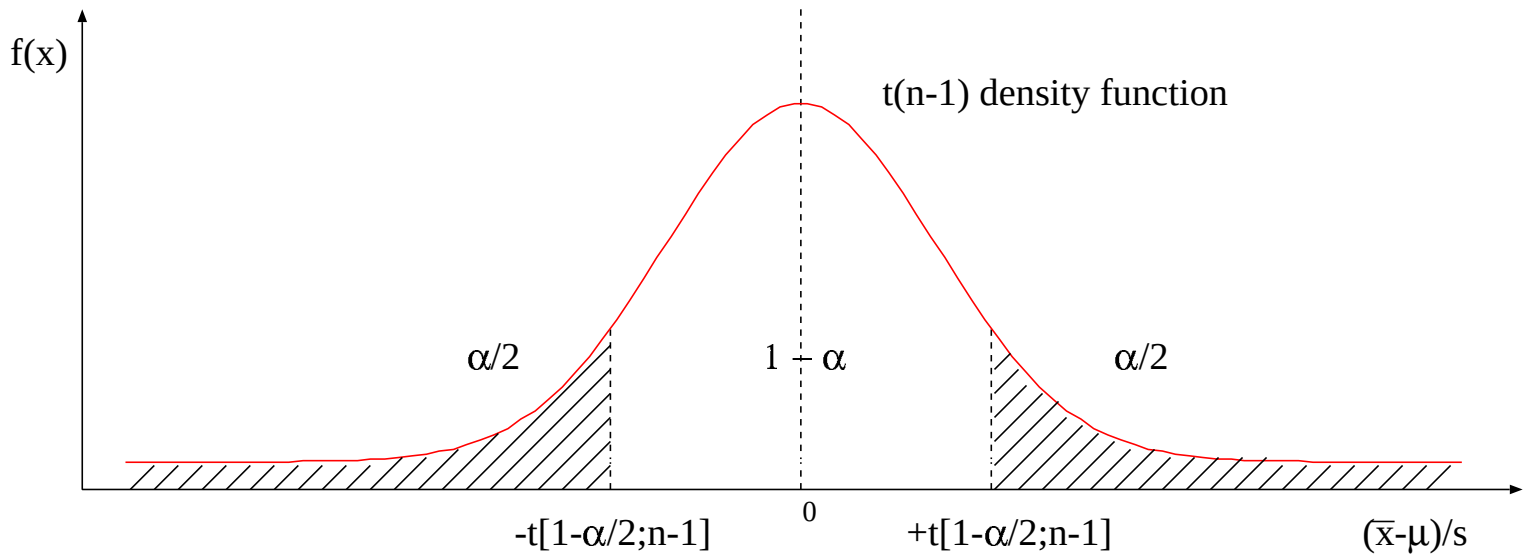
サンプル数が少ない場合の平均値の信頼区間

サンプル数が少ない (< 30) 場合、母集団が正規分布に従う場合に限って、信頼区間を求める事ができる

- 正規分布からサンプルを取った場合、標準誤差 $(\bar{x} - \mu)/(s/\sqrt{n})$ は $t(n-1)$ 分布となる

$$\bar{x} \mp t_{[1-\alpha/2;n-1]}\, s/\sqrt{n}$$

ここで、$t_{[1-\alpha/2;n-1]}$ は 自由度 $(n-1)$ の t 分布における $(1-\alpha/2)$ 領域の境界値



サンプル数が少ない場合の平均値の信頼区間の例

- 例: 前述の TCP スループット計測では、$t(n-1)$ 分布を使った信頼区間の計算をする必要
  - 95%信頼区間 $n = 5$: $t_{[1-0.05/2,4]} = 2.776$

    $$\bar{x} \mp 2.776(s/\sqrt{n}) = 3.56 \mp 2.776 \times 0.30/\sqrt{5} = 3.56 \mp 0.37$$

  - 90%信頼区間 $n = 5$: $t_{[1-0.10/2,4]} = 2.132$

    $$\bar{x} \mp 2.132(s/\sqrt{n}) = 3.56 \mp 2.132 \times 0.30/\sqrt{5} = 3.56 \mp 0.29$$

他の信頼区間

- 母分散:
  - 自由度 $(n-1)$ の χ^2 分布
- 標本分散の比:
  - 自由度 (n_1-1, n_2-1) の F 分布

信頼区間の応用

応用例

- 平均値の推定範囲を示す
- 平均と標準偏差から、必要な信頼区間を満足するために何回試行が必要か求める
- 必要な信頼区間を満足するまで計測を繰り返す

平均を得るために必要なサンプル数

- 信頼度 $100(1-\alpha)$ で $\pm r\%$ の精度で母平均を推定するためには何回の試行 n が必要か？
- 予備実験を行い 標本平均 $\bar{x}$ と 標準偏差 s を得る
- サンプルサイズ n、信頼区間 $\bar{x} \mp z\frac{s}{\sqrt{n}}$、必要な精度 $r\%$

$$\bar{x} \mp z\frac{s}{\sqrt{n}} = \bar{x}(1 \mp \frac{r}{100})$$

$$n = (\frac{100zs}{r\bar{x}})^2$$

- 例: TCP スループットの予備計測で、標本平均 3.56Mbps、標本標準偏差 0.30Mbps を得た。
  信頼度 95%、精度 (< 0.1Mbps) で平均を得るためには何回測定する必要があるか？

$$n = (\frac{100zs}{r\bar{x}})^2 = (\frac{100 \times 1.960 \times 0.30}{0.1/3.56 \times 100 \times 3.56})^2 = 34.6$$

推定と仮説検定

仮説検定 (hypothesis testing) の目的

- 母集団について仮定された命題を標本に基づいて検証

推定と仮説検定は裏表の関係

- 推定: ある範囲に入ることを予想
- 仮説検定: 仮説が採用されるか棄却されるか
  - 母集団に入るという仮説を立て、その仮説が 95%信頼区間に入るかを計算
  - 区間内であれば仮説は採用される
  - 区間外では仮説は棄却される

検定の例

N 枚のコインを投げて表が 10 枚でた。 この場合の N として 36 枚はあり得るか？ (ただし分布は $\mu = N/2, \sigma = \sqrt{n}/2$ の正規分布にしたがうものとする)

- 仮説: $N = 36$ で表が 10 枚出る
- 95%信頼度で検定

$$-1.96 \leq (\bar{x} - 18)/3 \leq 1.96 \quad 12.12 \leq \bar{x} \leq 23.88$$

10 は 95%区間の外側にあるので 95%信頼度では $N = 36$ という仮説は棄却される

課題 1 の解答

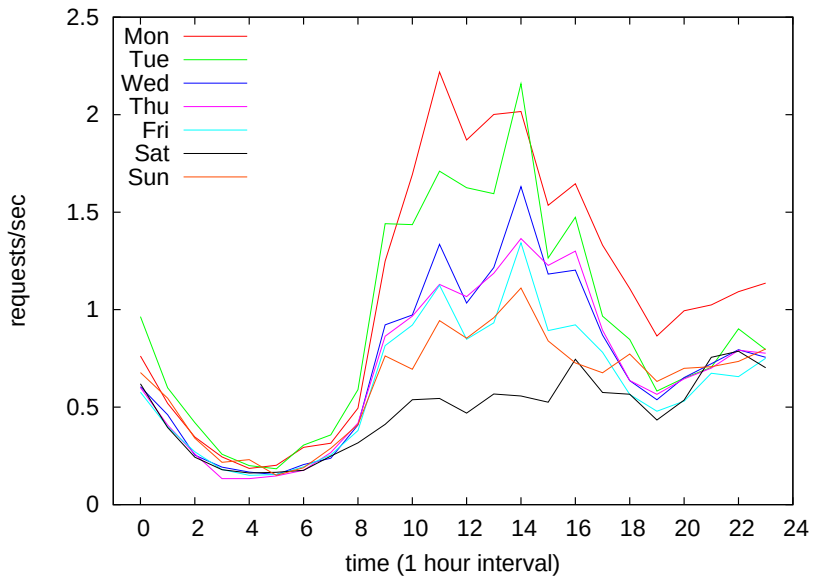
プロット用データテーブルの作成

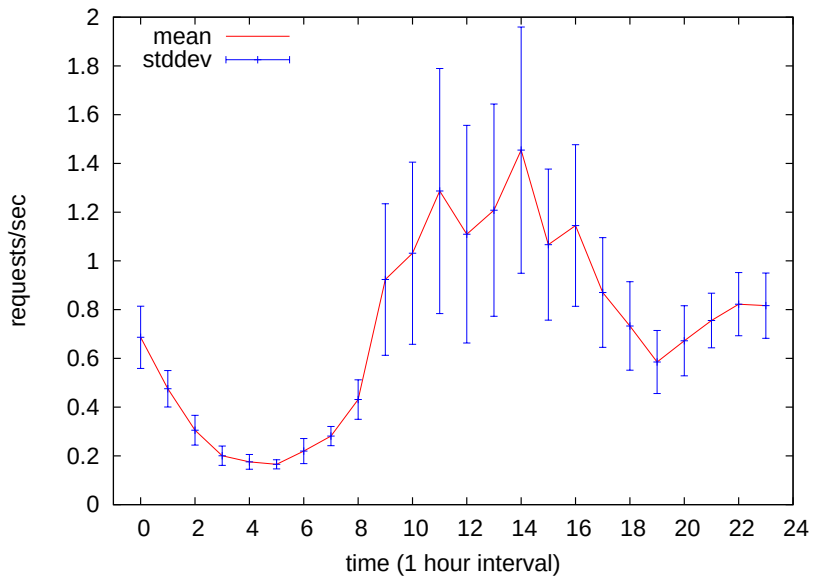
#hour	Mon	Tue	Wed	Thu	Fri	Sat	Sun	mean	stddev
0	2745	3472	2177	2161	2076	2230	2439	2471.4	459.00
1	1869	2157	1660	1469	1437	1423	1964	1711.3	269.15
2	1249	1515	914	935	975	877	1228	1099.0	219.73
3	880	931	693	481	649	644	781	722.7	142.66
4	668	720	601	481	541	585	831	632.4	108.82
5	728	662	553	530	557	597	541	595.4	68.17
6	1058	1100	739	633	699	633	681	791.9	185.15
7	1134	1288	858	960	917	896	1041	1013.4	141.82
8	1781	2122	1471	1507	1368	1141	1477	1552.4	291.38
9	4496	5187	3319	3110	2936	1482	2746	3325.1	1119.17
10	6093	5170	3502	3480	3315	1937	2499	3713.7	1345.49
11	7987	6157	4807	4066	4051	1960	3398	4632.3	1809.99
12	6732	5851	3721	3840	3054	1689	3073	3994.3	1607.49
13	7204	5740	4378	4272	3357	2043	3451	4349.3	1567.56
14	7259	7770	5871	4913	4834	2007	3999	5236.1	1818.14
15	5527	4554	4257	4417	3214	1891	3024	3840.6	1116.05
16	5924	5306	4330	4680	3319	2684	2616	4122.7	1194.29
17	4790	3479	3131	3221	2813	2070	2434	3134.0	810.12
18	3991	3050	2289	2291	2034	2038	2780	2639.0	653.29
19	3113	2097	1936	2036	1724	1562	2274	2106.0	465.68
20	3579	2338	2345	2323	1910	1930	2517	2420.3	517.50
21	3687	2535	2601	2520	2425	2721	2547	2719.4	403.63
22	3932	3245	2860	2850	2363	2833	2645	2961.1	465.96
23	4089	2864	2721	2795	2703	2528	2874	2939.1	481.85

課題 1 の解答 (つづき)

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





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 "± 1.960 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)

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

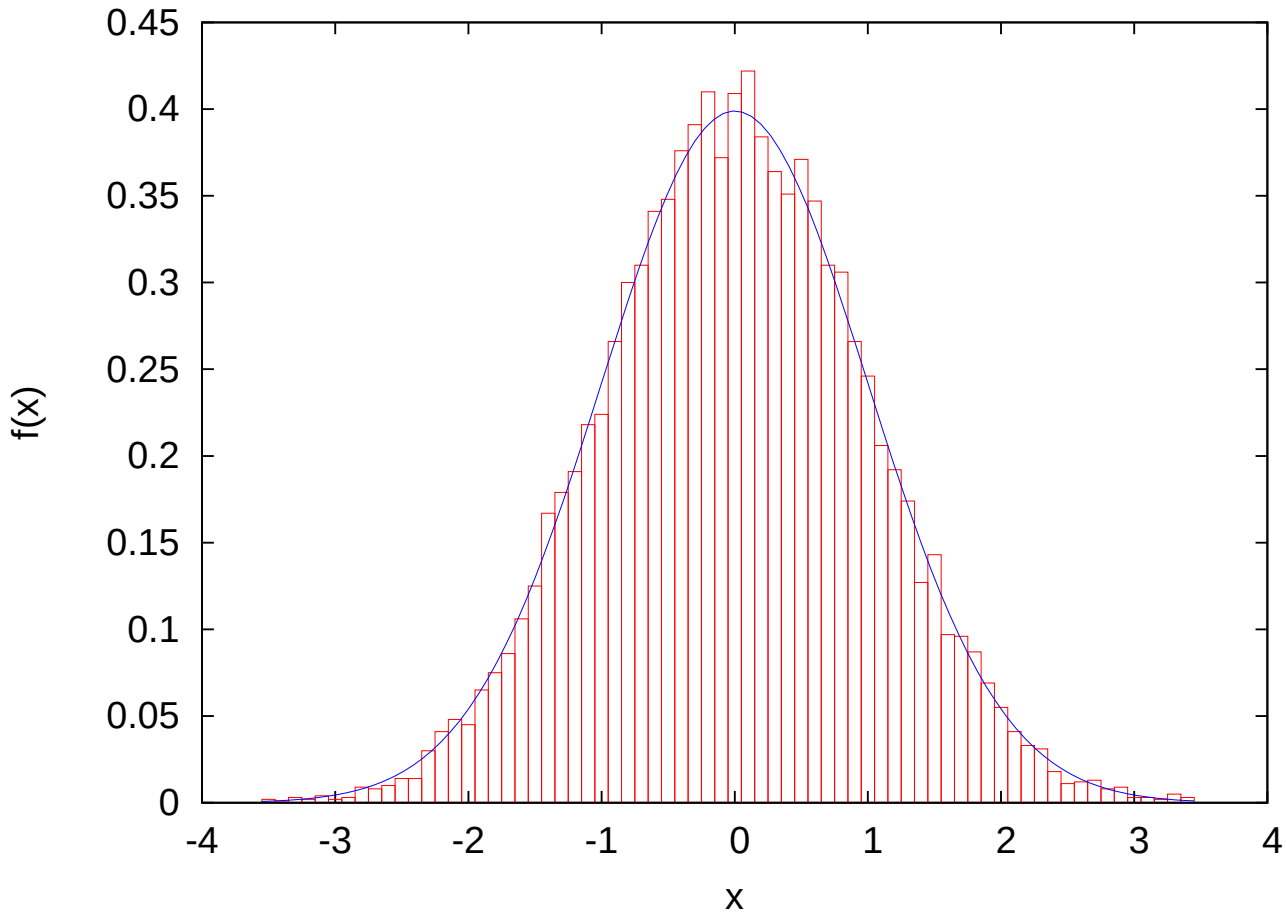
$$z_1 = u_1\sqrt{\frac{-2\ln s}{s}}$$

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

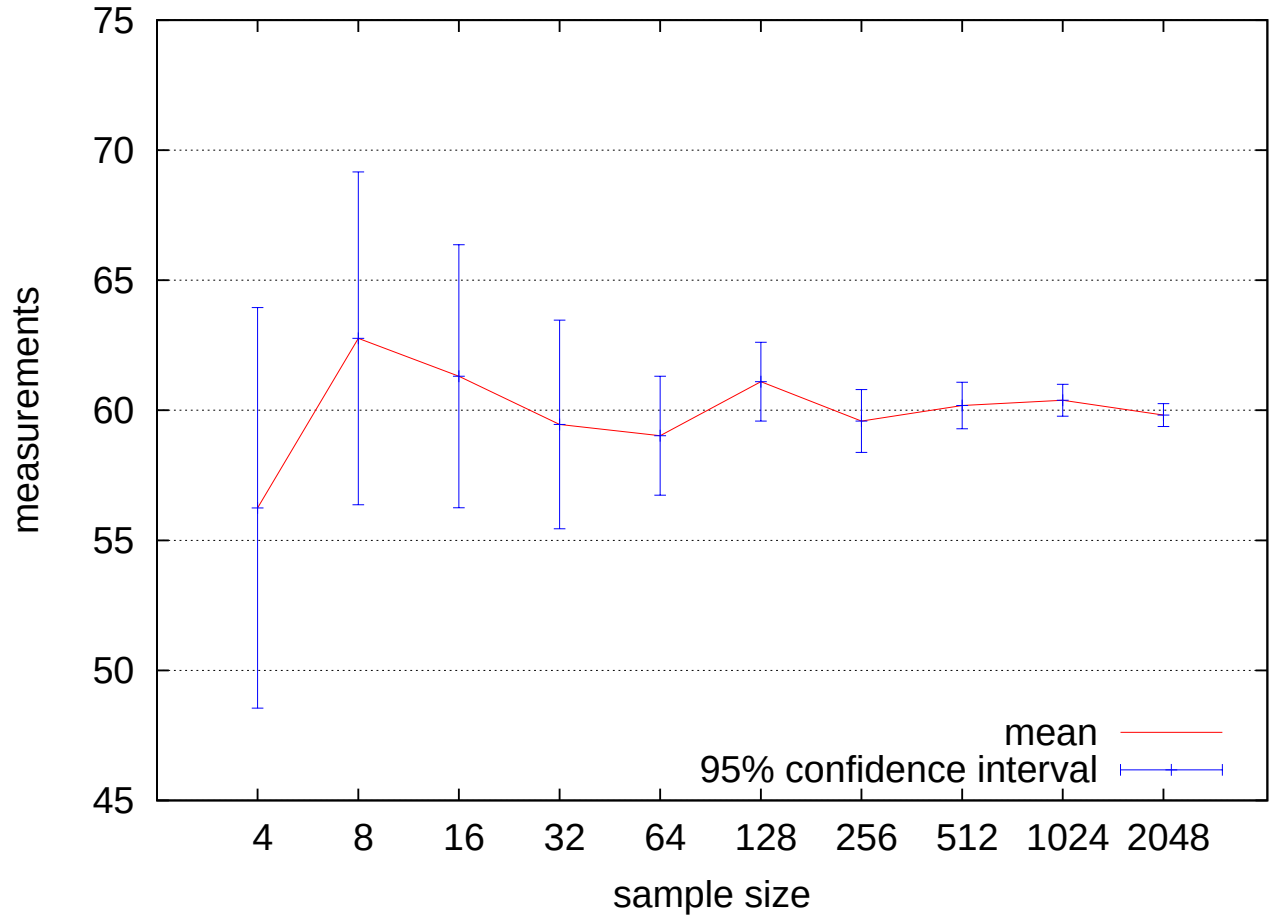
```
#
# create histogram: bins with 1 digit after the decimal point
#

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

まとめ

ロングテールとさまざまな分布

- 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

次回予定

第 9 回 インターネットの時間変化を計る (6/29)

- インターネットと時刻
- ネットワークタイムプロトコル
- 時系列解析
- 演習:時系列解析