インターネット計測とデータ解析 第6回

長 健二朗

2010年11月10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
- グラフによる可視化
- 相関と多変量解析

今日のテーマ

インターネットの多様性と複雑さを計る

- ロングテールとさまざまな分布
- サンプリング
- 統計解析 (期待値と大数の法則、信頼区間と検定)

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:次回授業のテーマ)

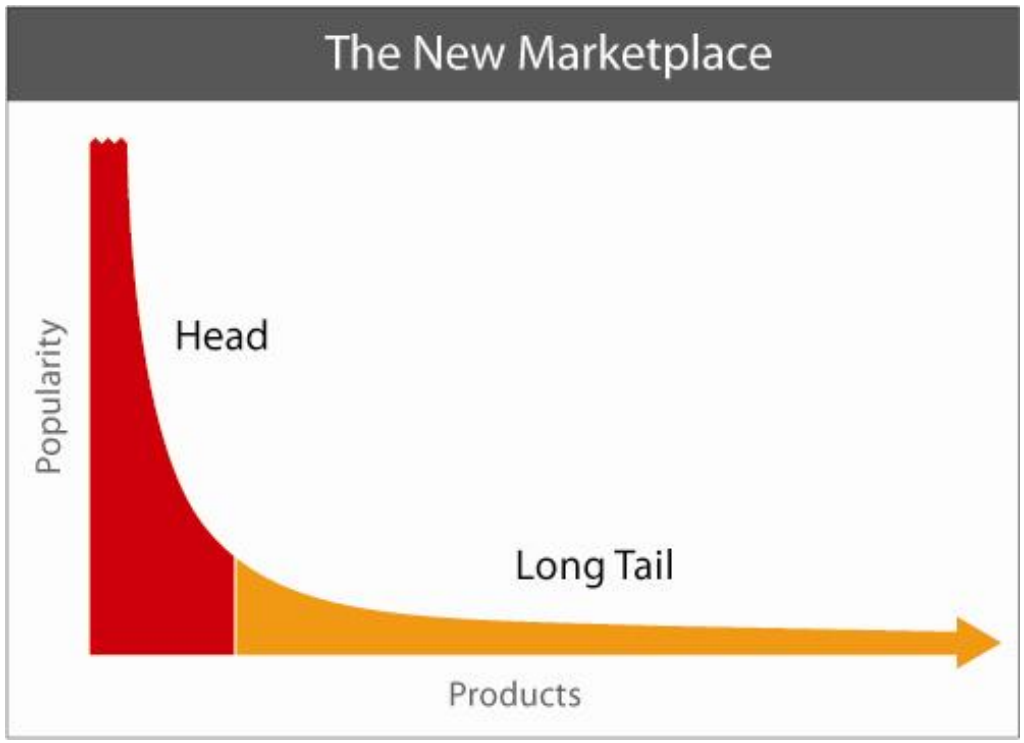
- 自己相似性
- 長期依存性

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/

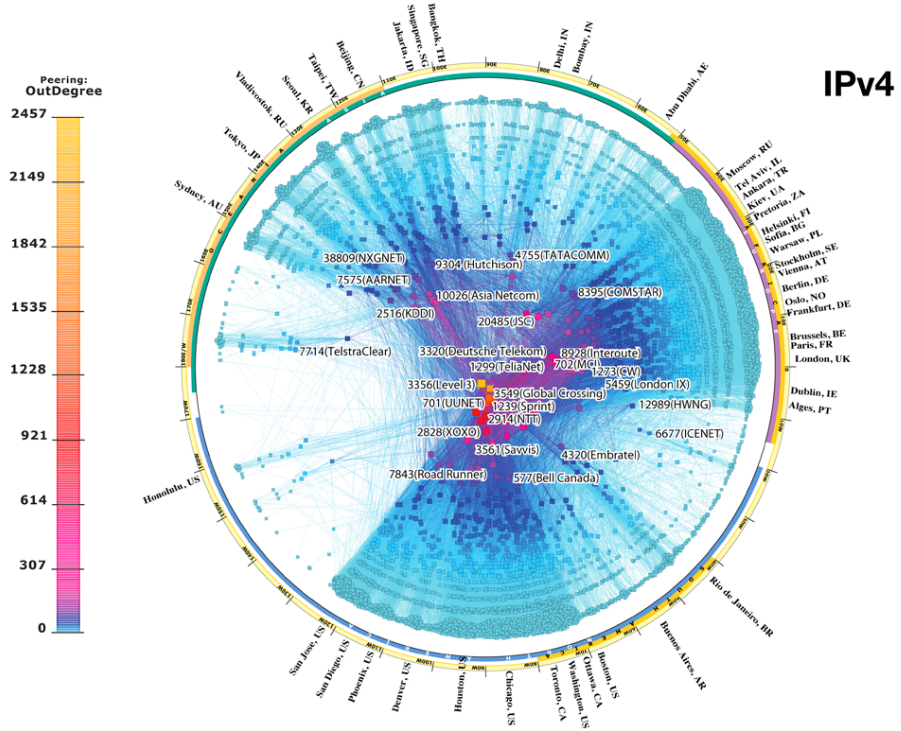
インターネットの AS 構造の例

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

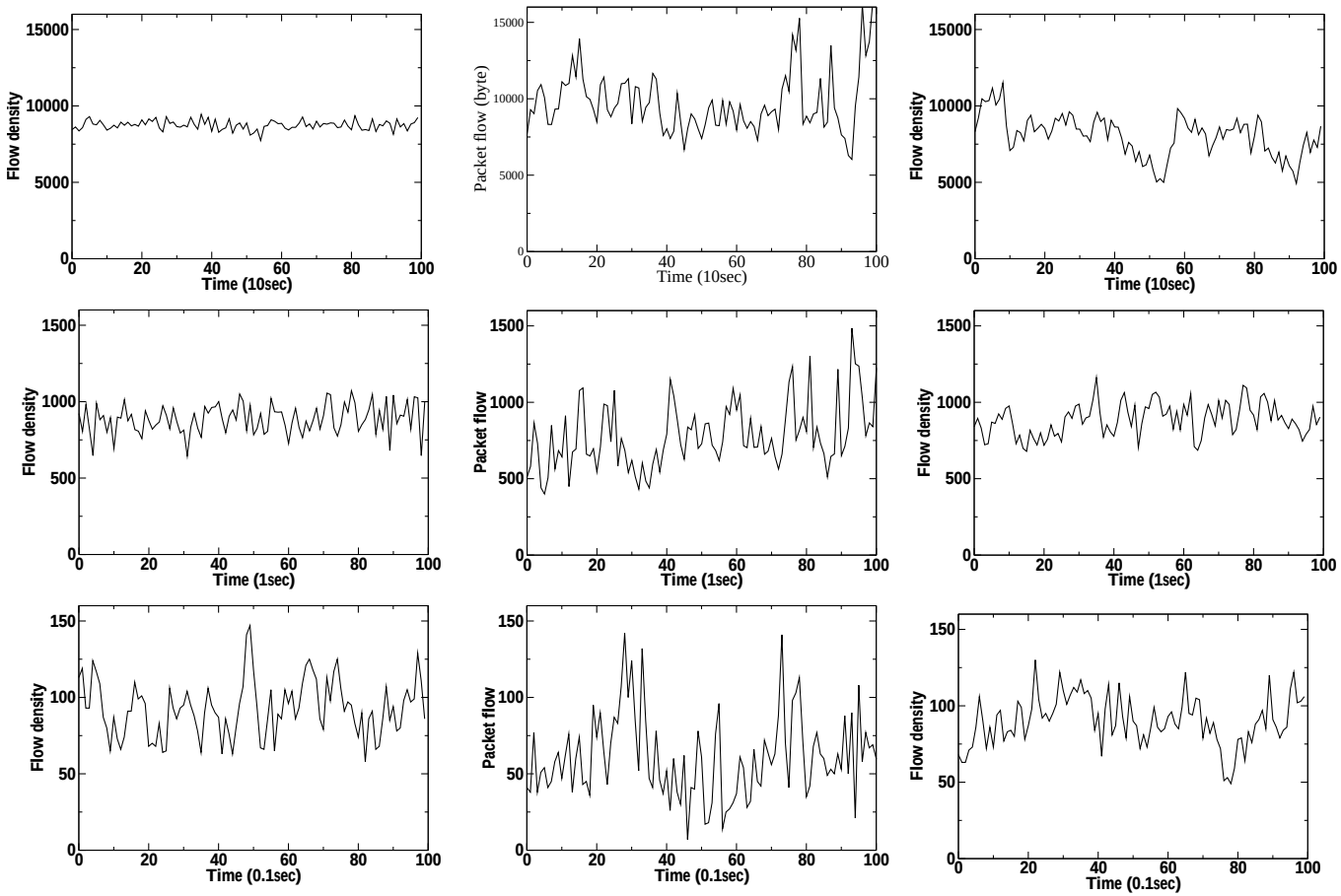
AS-level INTERNET GRAPH



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

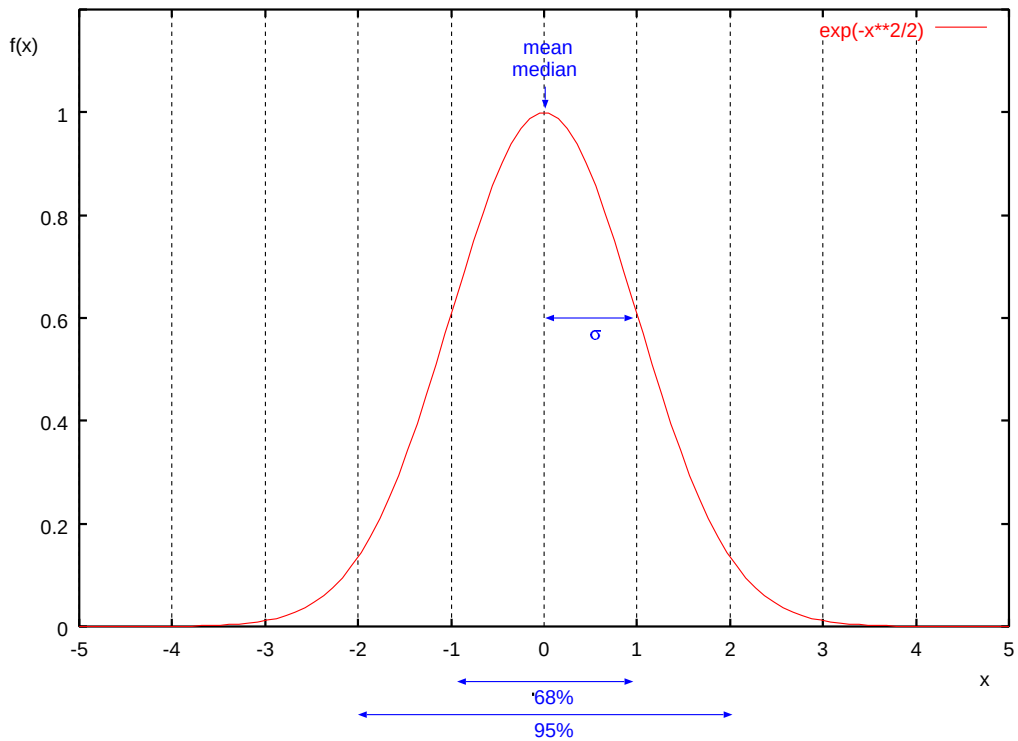


さまざまな分布

- 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

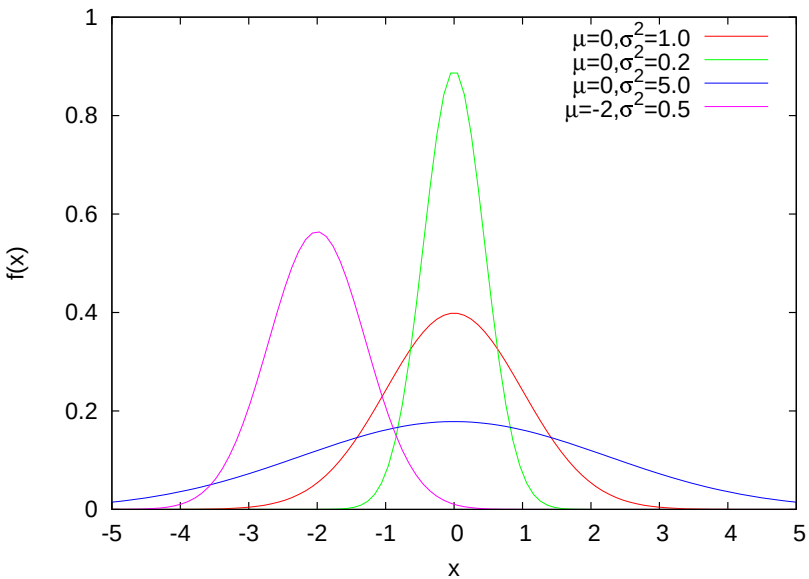
確率密度関数 (PDF)

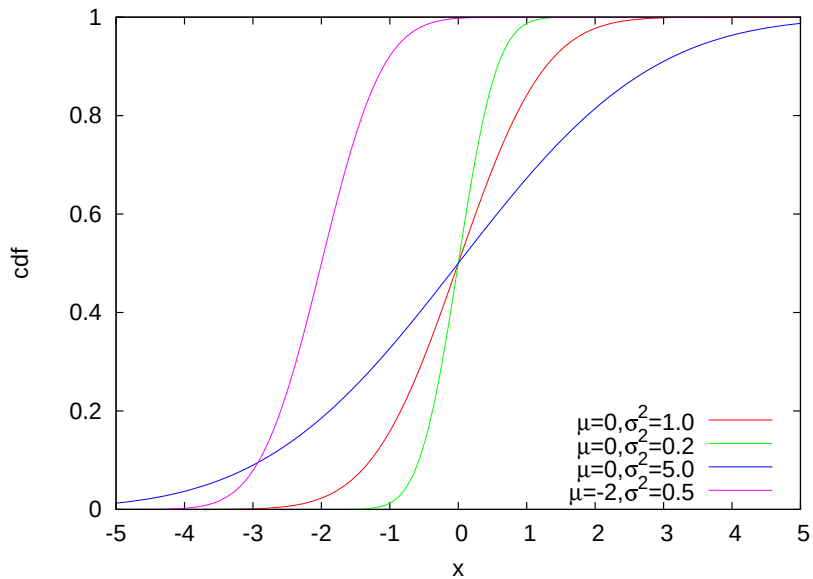
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

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)

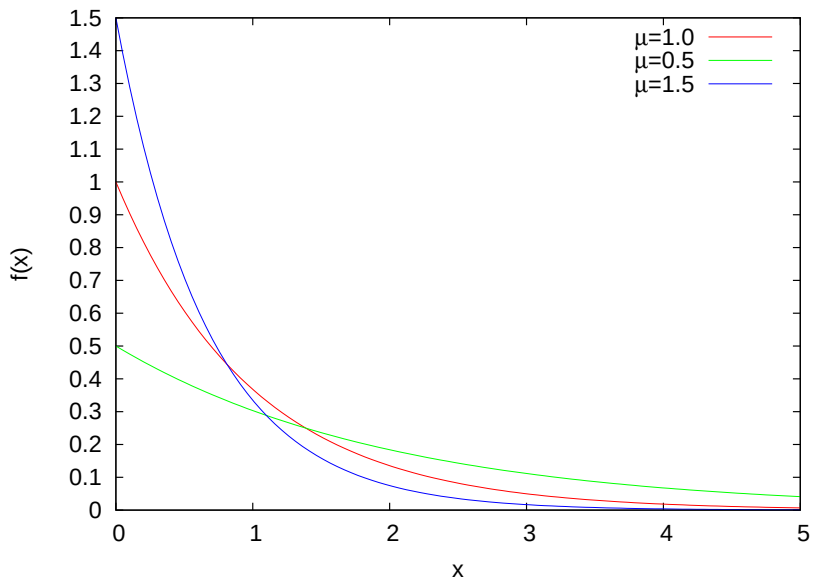
$$f(x) = \lambda e^{-\lambda x}, (x \geq 0)$$

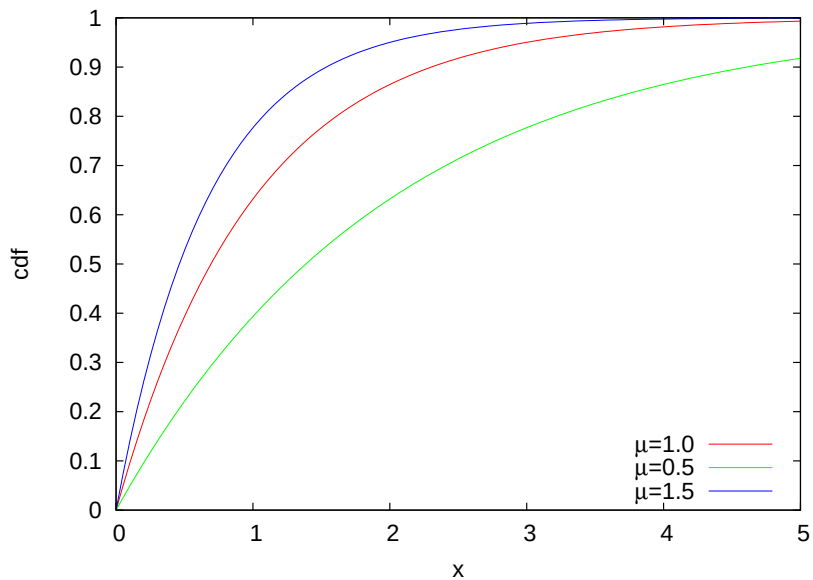
累積分布関数 (CDF)

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

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

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





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)

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

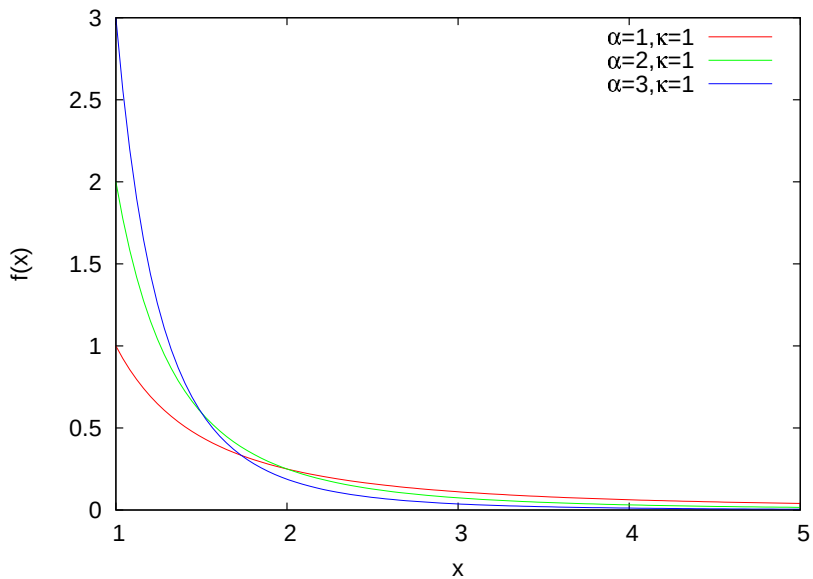
累積分布関数 (CDF)

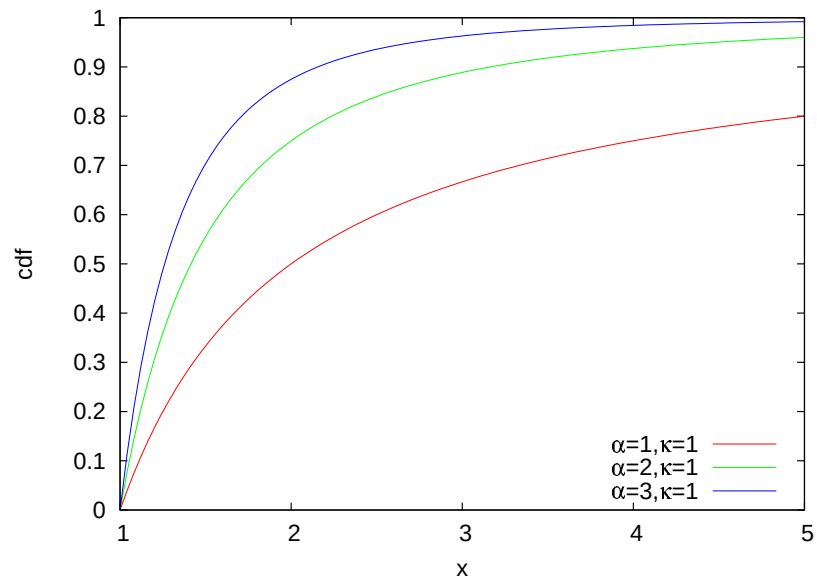
$$F(x) = 1 - (\frac{\kappa}{x})^{\alpha}$$

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

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

if $\alpha \leq 2$, variance $\to \infty$. 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

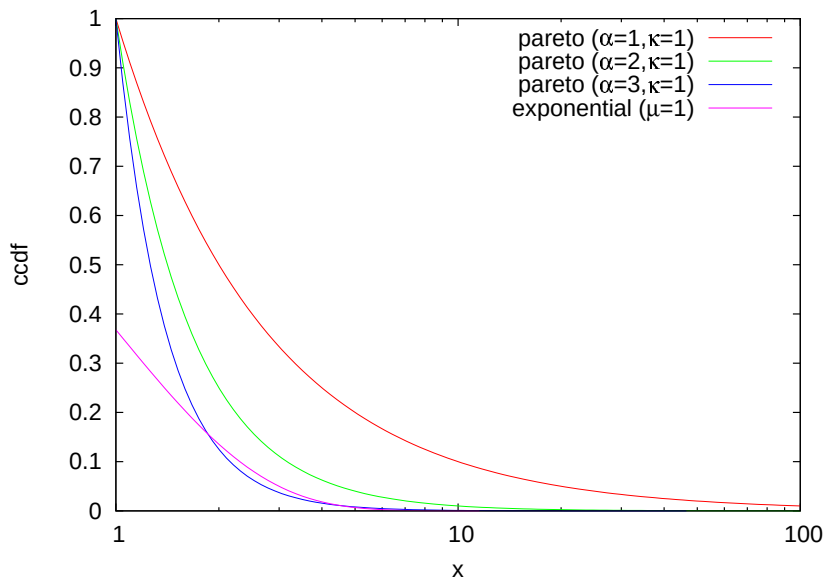
CCDF: x より大きい値の合計が全体に占める割合

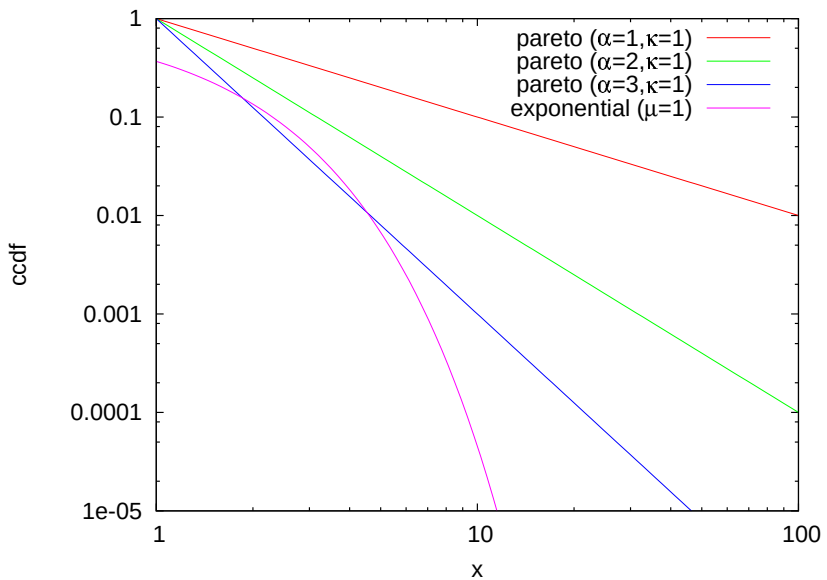
$$F(x) = 1 - P[X <= x]$$

- 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

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





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

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

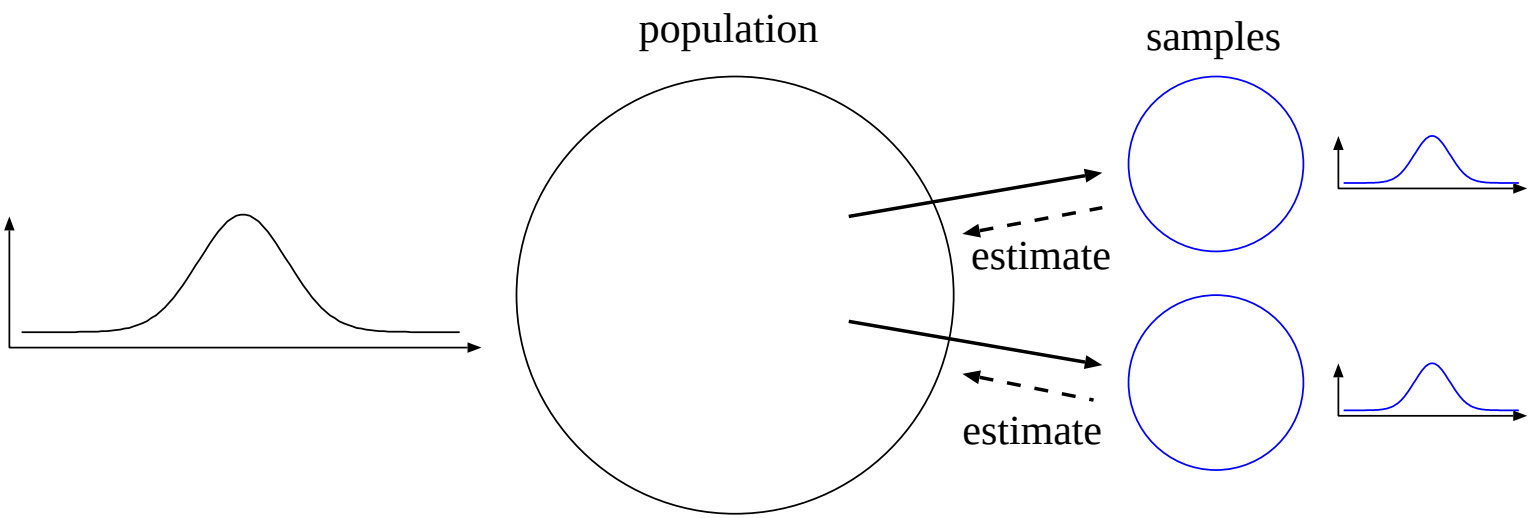
期待値の性質

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

サンプリング: 標本と母集団

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)



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: 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

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

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*

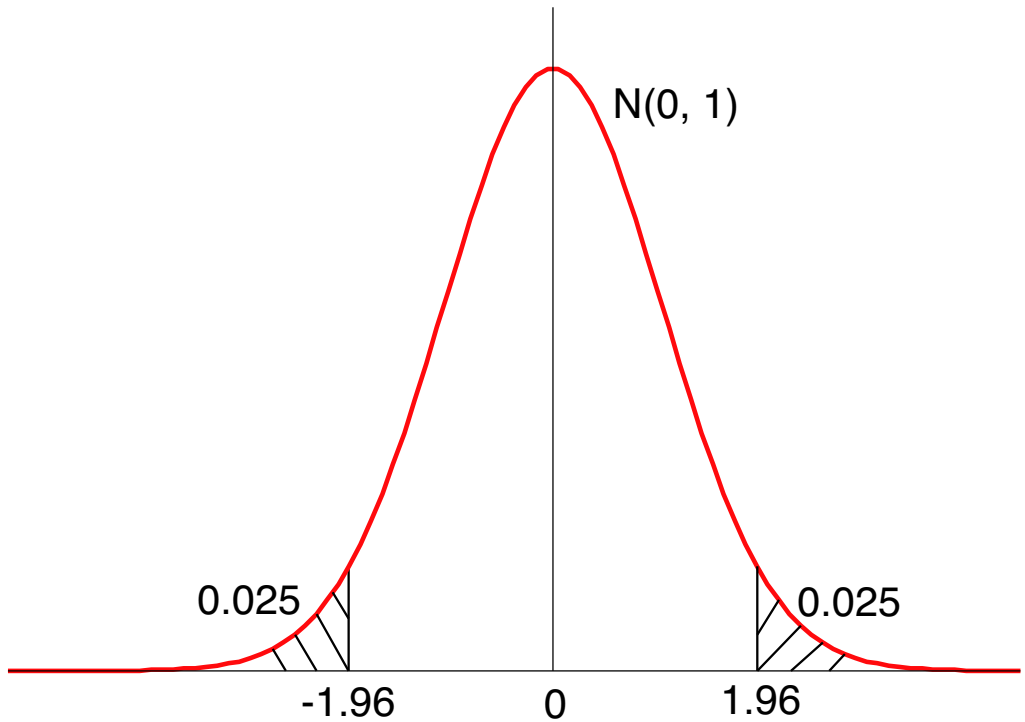
- 例: 信頼度 95% で、母平均は、$c1$ と $c2$ の間に存在
- 慣習として、信頼度 95% と 99% がよく使われる

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

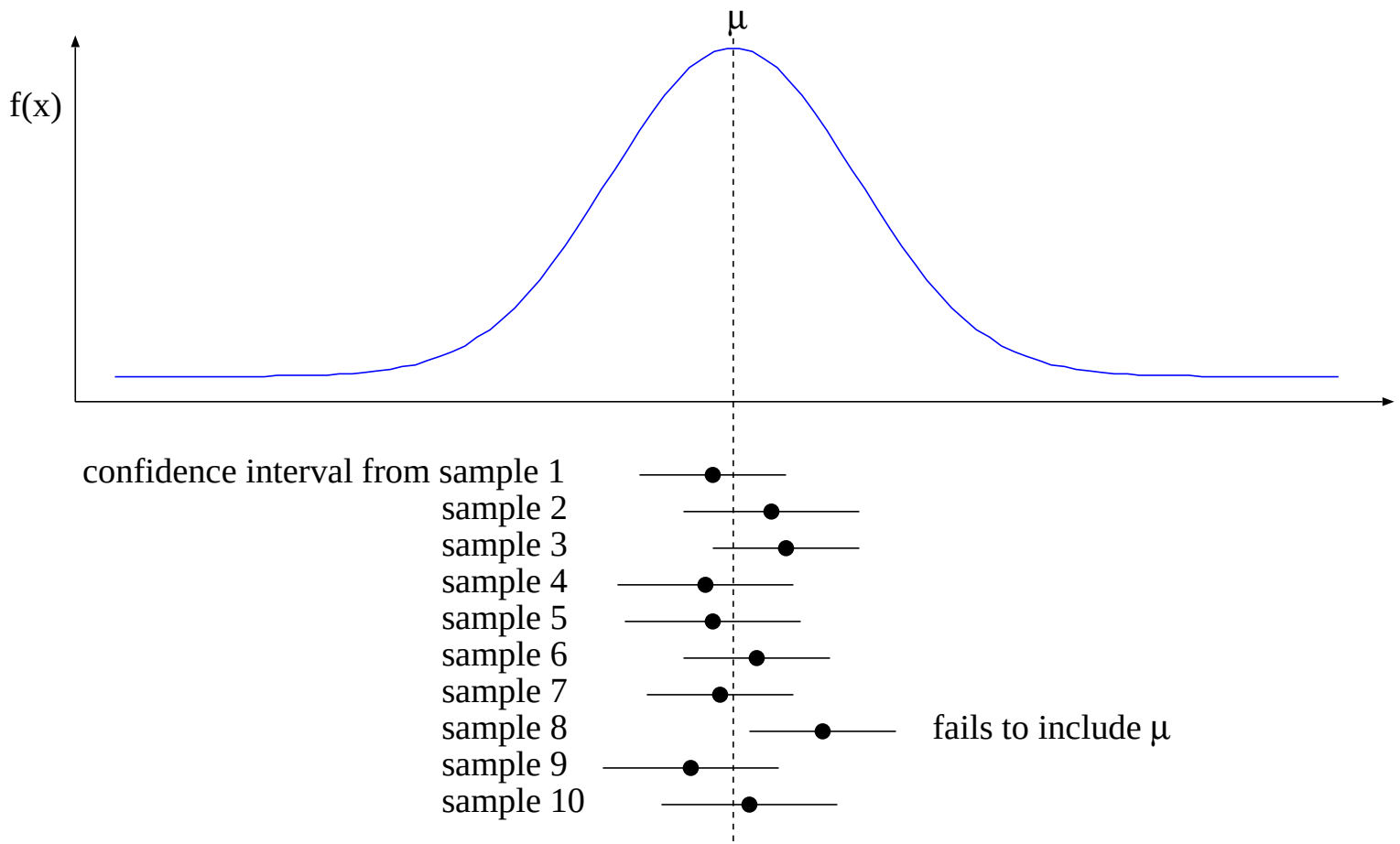
$$-1.96 \leq \frac{\bar{x} - \mu}{\sigma\sqrt{n}} \leq 1.96$$



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:

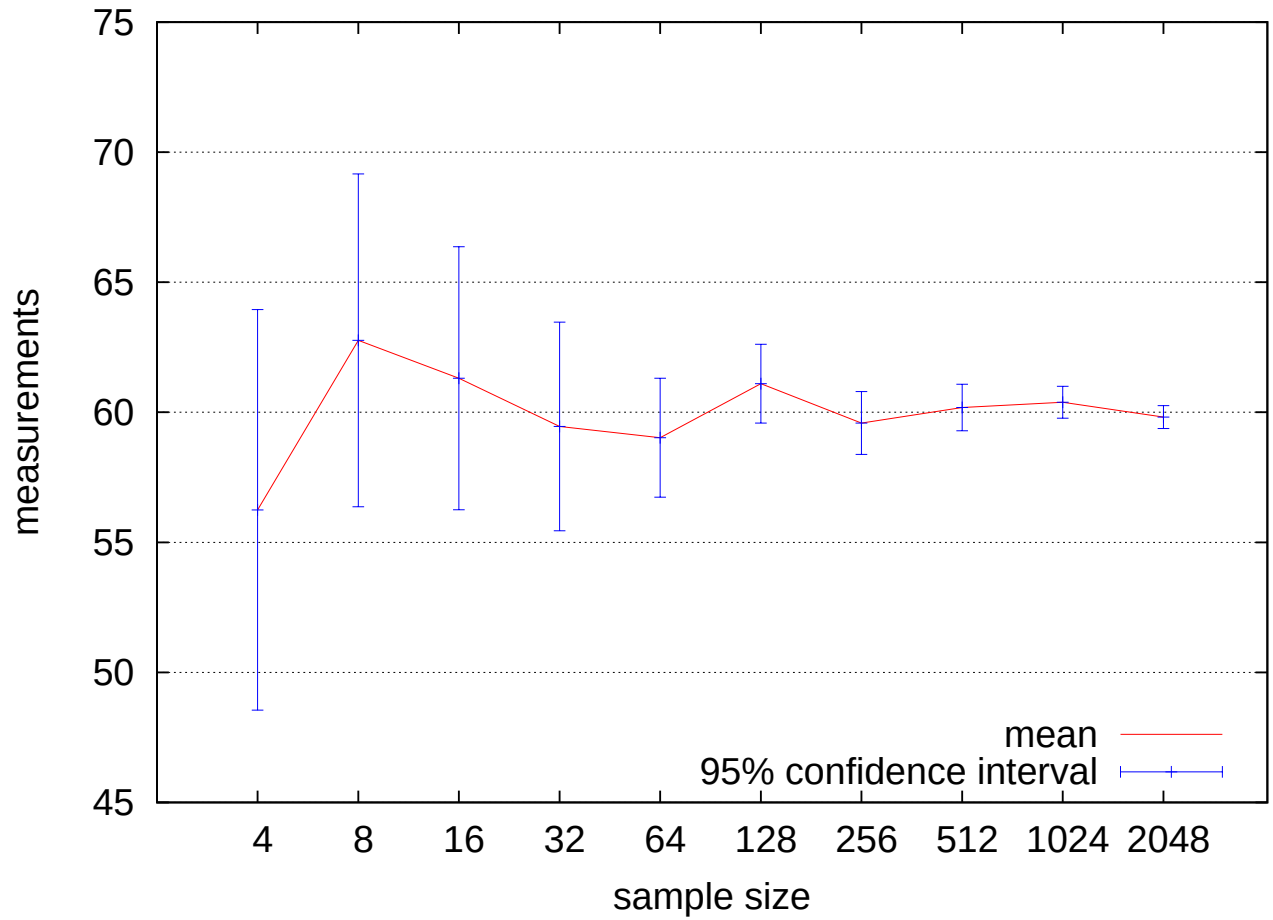
    $$\bar{x} \mp 1.96(s/\sqrt{n}) = 3.56 \mp 1.960 \times 0.30/\sqrt{5} = 3.56 \mp 0.26$$

  - 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

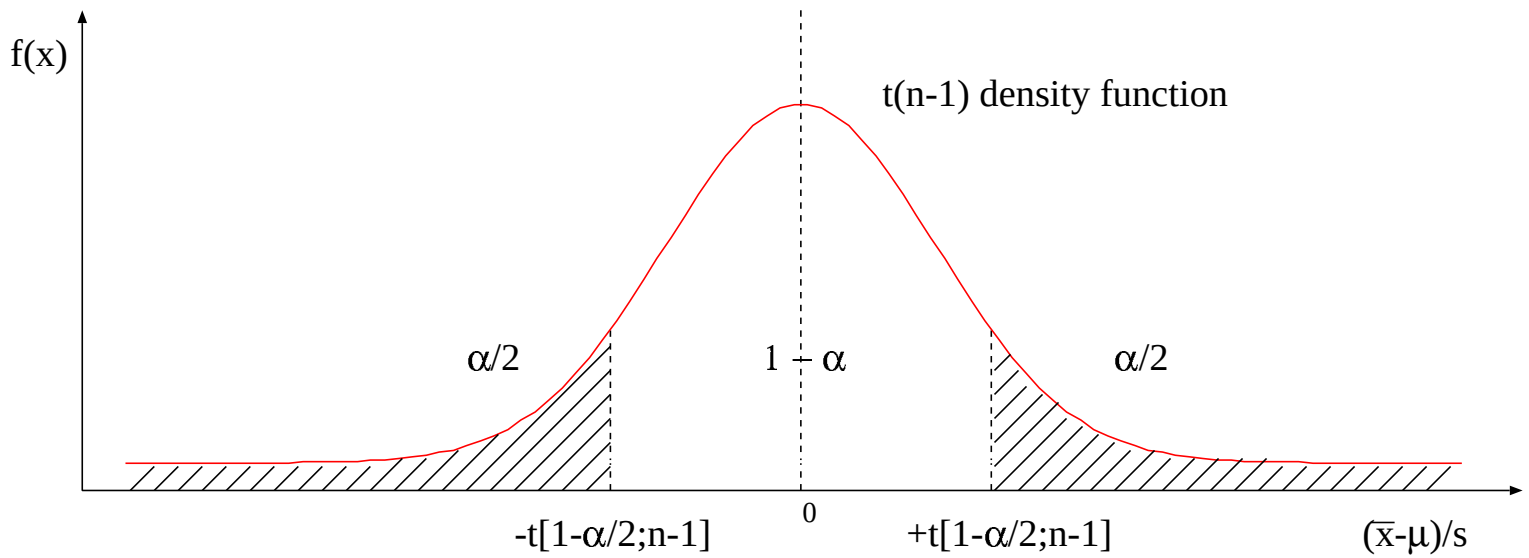
サンプル数が少ない場合の平均値の信頼区間

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

まとめ

インターネットの多様性と複雑さを計る

- ロングテールとさまざまな分布
- サンプリング
- 統計解析 (期待値と大数の法則、信頼区間と検定)

次回予定

第 7 回 インターネットの時間変化を計る (11/17)

- インターネットと時刻
- 時系列解析
- **課題 2**