Rubyによるデータ解析
Data Analysis in Ruby

2016年2月19日

株式会社ネットワーク応用通信研究所　前田　修吾

NaCl

データ解析

■ Data Analysis

Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making.

Data analysis
https://en.wikipedia.org/wiki/Data_analysis

■ 目的

- 有用な情報の発見、結論の提案、意思決定の支援

■ 手段

- データの検査、クリーニング、変換、モデル化

具体例

株式市場のブラウン運動

- M. F. M. Osborne, *Brownian Motion in the Stock Market*, Operations Research, 1959
- 株価の対数とブラウン運動における微粒子の座標との類似性
- 統計力学的手法を株価に適用
- 時系列データの分析

1. 株価の変化

- 株価の変化は離散的（1/8ドル単位）
- 株価の対数も同じ

2. 取引数

- **単位時間あたりに有限の取引（あるいは決定）が行われる**
  - 一つの株に対して0～1000あるいはそれ以上

3. Weber-Fechnerの法則

- **精神物理学の基本法則**
- **感覚量Eは刺激量の強度Rの対数に比例する**
  - $E = C \log R$
  - 強度100の刺激が200に増加した場合の感覚量
＝
強度200の刺激が400に増加した場合の感覚量
- **株価の刺激とそれに対するトレーダー・投資家の主観的感覚はこの法則に従うと仮定する**

統計学的アプローチ

- **金融の知識のない統計学者がNY市場の取引データを分析したら？**
  - 集団が均質かどうか
  - 各属性・変数の関連性

株価の分布

- 株価の終値を主要な変数と推測
- 1000要素のサンプルの分布をプロット

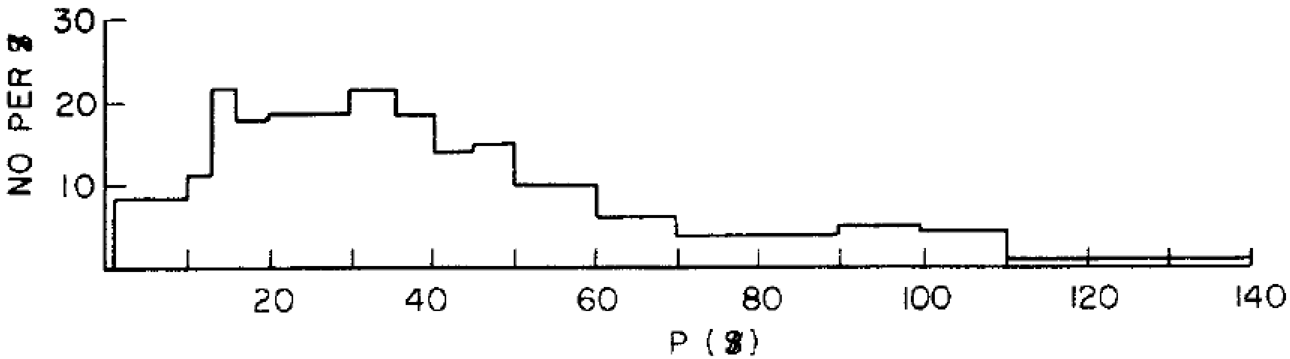


Fig 1 Distribution function of closing prices for July 31, 1956 (all items, NYSE)

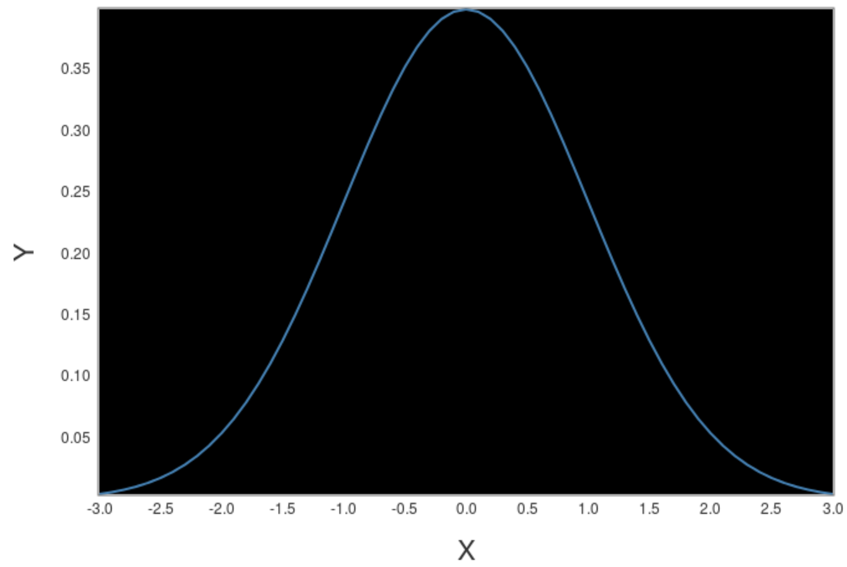
- 株価は正規分布に従わない
- 株価の対数は正規分布に従うかもしれない

平均・分散・標準偏差

- **平均**
  - $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **分散**
  - 平均からのばらつきの指標（正の値にするため自乗誤差を使う）
  - $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$
- **標準偏差**
  - 分散の平方根（x_iやμの値と比較しやすくするため平方根を取る）
  - $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$

正規分布

- **以下の確率密度関数を持つ**
  - $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- **確率密度関数**
  - 定積分（面積）が確率
- **計算で扱いやすい**
- **サンプル数を増やすと平均が収束（大数の法則）**
- **正規分布で近似できる対象が多い（中心極限定理）**
- **$\mu = 0, \sigma = 1$のとき、標準正規分布と呼ぶ**



株価の対数の分布

- **正規分布ではない**
- **$\log_e P \approx 45$周辺の副極大**
  - この集団は均質でない
  - 少なくとも二つの下位集団
- **生データの確認**
  - $\log_e P \approx 45$周辺のデータにpfd (preferred) 属性

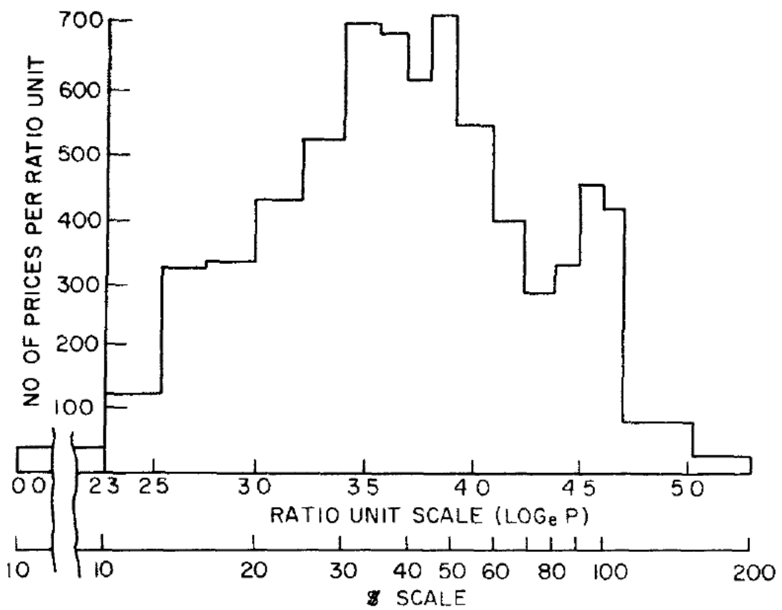


Fig 2 Distribution function for $\log_e P$ on July 31, 1956 (all items NYSE)

NaCl

対数を使う根拠

- 価格の変化と利益・損失に対する主観的感覚を「同じ間隔」で測る
- $10から$11の価格変化と、$100から$110の価格変化に対する主観的感覚は同じ

4. 論理的決定

収益の期待値

- Aという一連の行動がY_{A1}, Y_{A2}という収益を確率$\varphi(Y_{A1})$, $\varphi(Y_{A2})$で生む
- Bという一連の行動についても同様に考える
- 収益の期待値 $\varepsilon(Y_A) = \sum_i Y_{Ai}\varphi(Y_{Ai})$

収益の期待値が高い行動を選択

- $\varepsilon(Y_A)$ と$\varepsilon(Y_B)$ のどちらが大きいか？

価格$\boldsymbol{P_0(t)}$の株を100株買うかどうか？

- $A =$ 将来 $t+\tau$ に株を売るために買う
- $\mathrm{B} =$ 買わない
- $Y_A(\tau) = \Delta\log_e[100\ P(t)] = \log_e[P(t+\tau)/P_0(t)]$
- $Y_B = 0$
- $Y_A(\tau)$の期待値の見積が正か負かによって論理決定を行う

5. 市場の平等性

- **$\Delta\log_e P$ を買い手は正、売り手は負と判断する**
  - $\mathrm{E}\varepsilon(\Delta\log_e P)_{\mathrm{S}} + \mathrm{E}\varepsilon(\Delta\log_e P)_{\mathrm{B}} = 0$
    - ここで P は1株当たりの価格、$\mathrm{E}\varepsilon$ は期待値の見積である
- **市場全体では以下の式のような状況**
  - $\mathrm{E}\varepsilon(\Delta\log_e P)_{\mathrm{M=S+B}} = 0$
  - 上記の式では見積を表す E はなくてもよいかもしれない

6. 株価の収益率の分布

- **以下の $Y(\tau)$ は、平均 0、標準偏差 $\boldsymbol{\sigma}_{Y(\tau)}$ の正規分布に従うと予測される**
  - $Y(\tau) = \log_e[P(t+\tau)/P_0(t)]$
- **$\boldsymbol{\sigma}_{Y(\tau)}$ は取引数の平方根に比例する**
- **取引数が時間上均一に分布すると考えると**
  - $\boldsymbol{\sigma}_{Y(\tau)}$ は時間間隔の平方根に比例する
  - すなわち、$\boldsymbol{\sigma}_{Y(\tau)}$ は $\sigma\sqrt{\tau}$ という形式となる

7. 数学的表現

- **$\boldsymbol{k}$ 個のランダムな独立変数 $\boldsymbol{y(i) = i, \cdots, k}$ を仮定する**
  - $y(i) = \Delta_{i\delta}\log_e P = \log_e[P(t+i\delta)/P(t+\{i-1\}\delta)]$
    - ここで、$P(t)$ はある銘柄の時間 t における価格、δ は取引間の小さな時間間隔である
    - $y(i)$ は同じ標準偏差$\sigma(i) = \sigma'$ を持つと仮定する
- **$\boldsymbol{k}$ 回の取引、$\tau = k\delta$ 時間後の $Y(\tau)$ を以下のように定義する**
  - $Y(\tau) = Y(k\delta) = \sum_{i=1}^{i=k} y(i) = \log_e[P(t+\tau)/P(t)] = \Delta_\tau \log_e P(t)$
- **Yの標準偏差**
  - $\boldsymbol{\sigma}_{Y(\tau)} = \sqrt{\varepsilon(Y^2) - [\varepsilon(Y)]^2} = \sqrt{\sum_{i=1}^{i=k} \sigma^2(i)} = \sqrt{k}\sigma' = \sqrt{\tau / \delta}\sigma'$

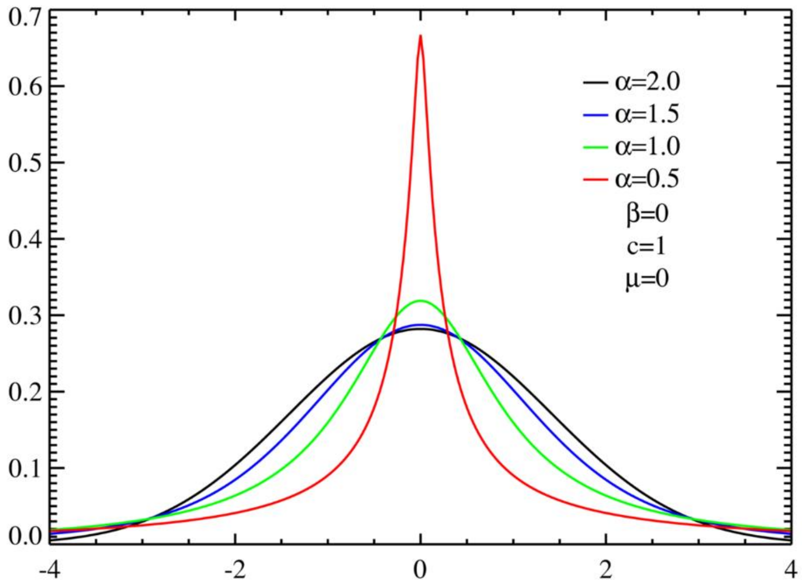
観測データとの比較

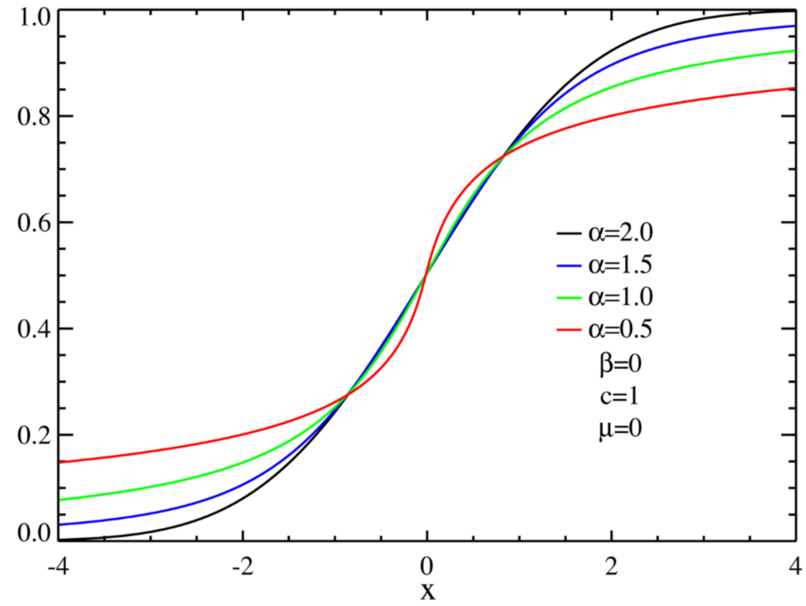
- 論文のデータについては省略
- 後でRubyでやってみる

収益率の分布の非正規性

- **Benoit Mandelbrot, *The Variation of Certain Speculative Prices*, Jounal of Business, 1963**
- **Benoit Mandelbrot, *The Variation of Other Speculative Prices*, Jounal of Business, 1967**
- **収益率の分布は、安定分布だが正規分布ではない**
- **α というパラメータで分布の裾の広さが決まる**
  - 小さいほど裾が広い（正規分布は$\alpha = 2$ ）
- **$\alpha \leq 1$ の場合、大数の法則に従わない**
- **$\alpha < 2$ の場合、分散が一定にならず中心極限定理が成り立たない**
  - Mandelbrotは収益率はこのケースの安定分布に従うと考えた

安定分布のPDFとCDF





左: 安定分布の確率密度関数

https://commons.wikimedia.org/wiki/File:Levy_distributionPDF.png public domain by PAR

右: 安定分布の累積分布関数

https://commons.wikimedia.org/wiki/File:Levy_distributionCDF.png CC-BY-SA 3.0 by PAR

Rubyによるデータ解析の現状

Why Ruby?

- **Pythonと同じ理由**
  - 「糊（グルー）」としてのRuby
    - C/C++/FORTRANなどで書かれたコードをつなぎ合わせる
  - 「2つの言語を利用する」ことの問題を解決する
    - アプリケーション開発とデータ解析で同じ言語を使う
- **でも道具は揃ってる？**

ライブラリ・ツール群

分類	Python	Ruby
ベクトル・行列	NumPy	NMatrix, NArray
データ解析	pandas	daru
科学計算	SciPy	SciRuby
可視化	matplotlib	Nyaplot
対話環境	Jupyter/IPython	Jupyter/IRuby

NMatrix vs NArray

daru

- Data Analysis in RUby

Daru::Vector

- 1次元ベクトル

```
v = Daru::Vector.new([40, 20, 30])
v.mean
```

Daru::DataFrame

- 2次元のスプレッドシート風データ構造

```
df = Daru::DataFrame.from_csv("n225.csv")
df.describe
```

Jupyter/IRuby

- **Jupyter**
  - Webアプリケーションによる対話環境
  - Matematica風のノートブック
    - プログラムの対話的実行
    - グラフの描画
  - プログラミング言語非依存
    - 各言語の実行環境をカーネルとして提供
    - プロセス間通信
- **IRuby**
  - JupyterのRubyカーネル

デモ

今後の課題

機能追加・機能改善

- **daru**
  - 欠損値の扱い
    - チェックや穴埋め
  - 時系列データの扱い
    - 再サンプリング
  - 統計量の計算
    - 共分散や相関係数
  - 金融関係の機能

性能改善

- **NMatrix**
  - 処理によって遅い？
- **daru**
  - GSLやNMatrixが利用できる場合はデフォルトで使う
  - ベクトル演算関数を暗黙的に利用する

DSLの強化

- **daru**
  - Daru::DataFrame#where
- **SciRuby**
  - 関数をRubyの式で表現

Refinementsによる拡張

計算と可視化の統合

- **daruはNyaplotを使用しているが、statsampleなどは使用していない**
- **計測データと理論値の重ね合わせが面倒**

References

- ジェイムズ・オーウェン・ウェザーオール, *ウォール街の物理学者*, 早川書房, 2015
- M. F. M. Osborne, *Brownian Motion in the Stock Market*, Operations Research, 1959
- Benoit Mandelbrot, *The Variation of Certain Speculative Prices*, Jounal of Business, 1963
- Benoit Mandelbrot, *The Variation of Other Speculative Prices*, Jounal of Business, 1967
- 平岡和幸・堀玄, *プログラミングのための確率統計*, オーム社, 2009
- Wes McKinnery, *Pythonによるデータ分析入門*, オライリー・ジャパン, 2013