統計学と星を観に行く

場所：１号館１５階　村上研究室

適合度検定

多くの統計的推論は母集団がある特定の分布に従うという仮定に基づいている．適合度検定は，未知な母集団分布$F(x)$と理論的な分布$F_0(x)$の同等性

帰無仮説 $H_0 : F(x) = F_0(x)$　vs.　対立仮説 $H_1 : F(x) \neq F_0(x)$

に対する仮説検定である．いくつかの有名な検定統計量として

$$\mathrm{KS} := \sup_{-\infty<x<\infty} |F_n(x) - F_0(x)| = \max\left\{ \max_{0\le i\le n}\left[\frac{i}{n} - F_0(X_{(i)})\right], \max_{1\le i\le n+1}\left[F_0(X_{(i)}) - \frac{i-1}{n}\right]\right\}, \quad \text{(Kolmogorov-Smirnov検定)}$$

$$\mathrm{CvM} := \int_{-\infty}^{\infty} (F_n(x) - F_0(x))^2 \, dF_0(x) = \frac{1}{12n^2} + \frac{1}{n}\sum_{i=1}^{n}\left[F_0(X_{(i)}) - \frac{i}{n} + \frac{1}{2n}\right]^2, \quad \text{(Cramer-von Mises検定)}$$

$$\mathrm{AD} := \int_{-\infty}^{\infty} \frac{(F_n(x) - F_0(x))^2}{F_0(x)(1-F_0(x))} dF_0(x) = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\log F_0(X_{(i)}) + \log\left(1 - F_0(X_{(n-i+1)})\right)\right] \quad \text{(Anderson-Darling検定)}$$

などがある． 特に，$F_0(x)$を正規分布とした正規性の検定が良く用いられる．

経験分布関数$F_n(x)$

$F(x)$からのランダム標本$X_1, X_2, \ldots, X_n$に対して，

$$F_n(x) = \frac{i}{n} \quad (X_{(i)} \le x < X_{(i+1)}).$$

ただし，$X_{(1)} \le X_{(2)} \cdots \le X_{(n)}$は順序統計量である．

※ 便宜上，$X_{(0)} = -\infty$，$X_{(n+1)} = \infty$と定義する．

MWGの等級データ

天の川

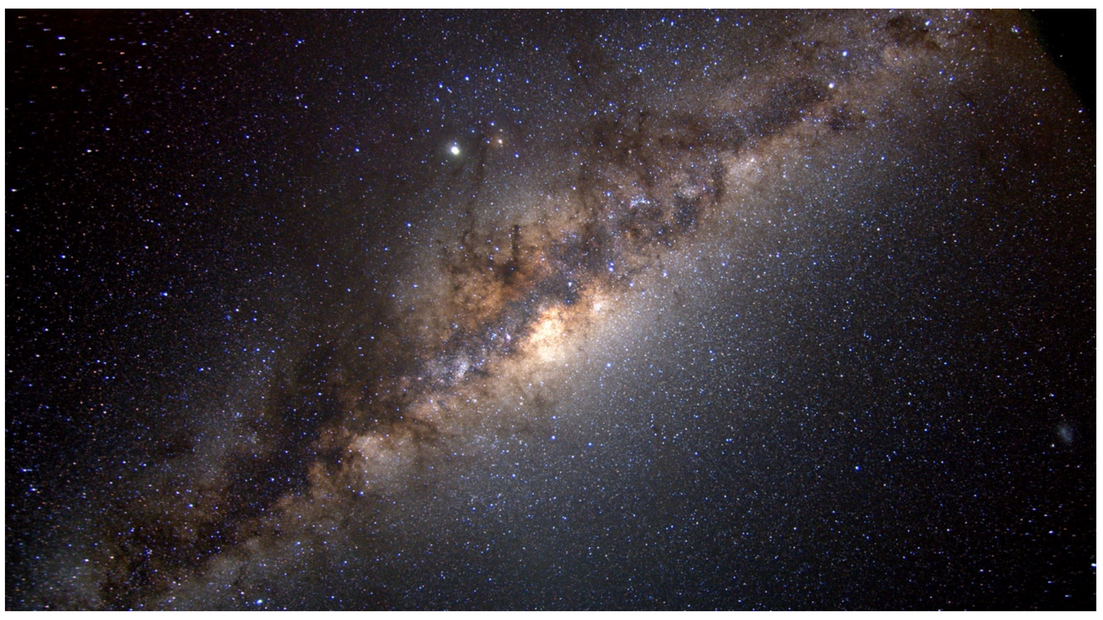
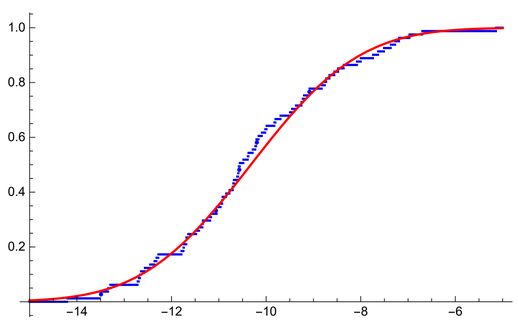


表2は，MWGの等級データ(絶対等級)に対して様々な検定統計量を用いて正規性の検定を行った結果である．

表2. 正規性の検定(MWG)

検定	統計量	p値
KS	0.0688	0.4498
CvM	0.0528	0.4662
AD	0.3026	0.5674

図4. MWGの経験分布と正規分布



〜〜〜考察〜〜〜

表2で示された各検定のp値は大きいため，MWGの等級データが正規分布に従うという帰無仮説を棄却できない．

また，図4のプロットからも経験分布関数と正規分布に大きな違いが見られないことが分かる．

仮に，MWGの等級データが正規分布に従っているとするならば，等級は天体の明るさの対数によって定められているので，MWGの球状星団の光度分布は対数正規分布に従うことが分かる．

M31の等級データ

M31の中心核

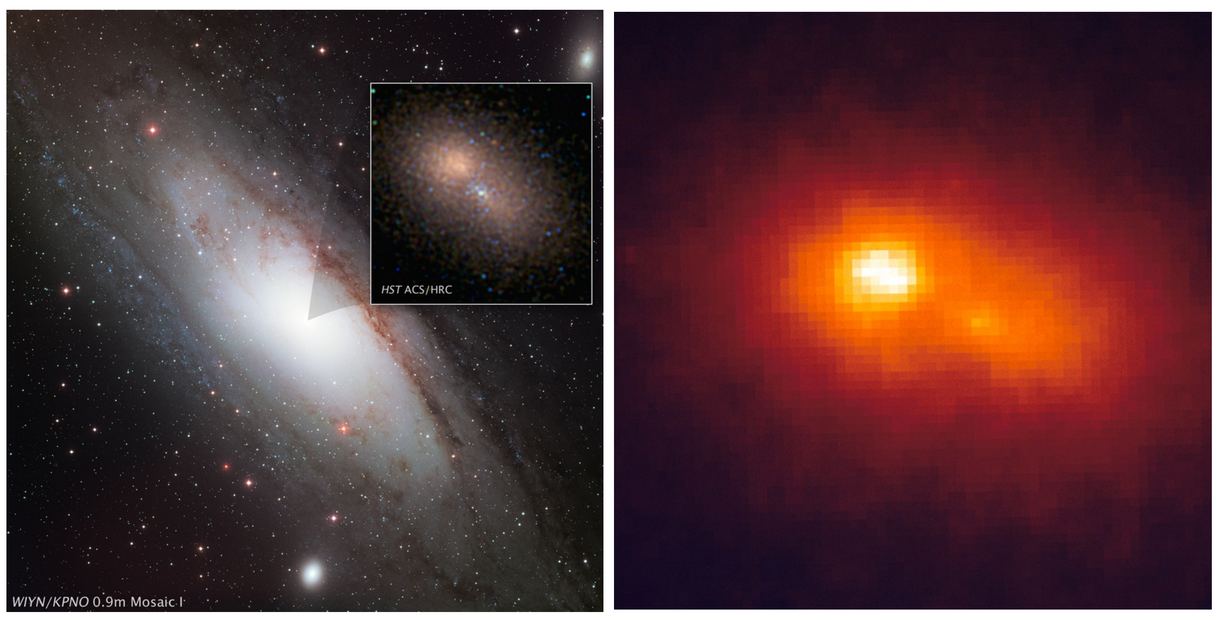
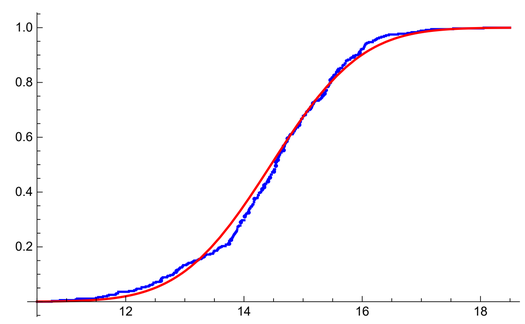


表3は，M31の等級データ(見かけの等級)に対して様々な検定統計量を用いて正規性の検定を行った結果である．

表3. 正規性の検定(M31)

検定	統計量	p値
KS	0.0634	0.0014
CvM	0.2894	0.0004
AD	1.7939	0.0001

図5. M31の経験分布と正規分布



〜〜〜考察〜〜〜

図5のプロットでは，M31の等級データに対する経験分布関数と正規分布には明らかな差はないように思われる．

しかし，表3で示された各検定のp値が非常に小さいため，M31の等級データが正規分布に従うという帰無仮説は棄却される．

この結果は，M31が2つの中心核をもつという構造に由来する可能性がある．従って，M31の等級データは2つの成分をもつ混合分布としてモデリングを行うのが適切であると予想される．

結果と今後の課題

☆ 天の川銀河(MWG)の等級が正規分布に従っていることを否定することはできない．

☆ MWGが正規分布に従うならば，等級の定め方が対数的であることから，MWGの球状星団の天体の光度は対数正規分布に従う．

☆ アンドロメダ銀河(M31)の等級は正規分布に従っているとは言えない．

☆ M31が2つの中心核を有するという構造を考慮して，混合分布モデルを用いることが適切であると予想される．

☆ 天体の観測技術の向上と，光度分布をより詳細に推定することによって，遥か遠方の銀河の構造が見えてくる可能性がある．

「ビッグデータの贈物」 ～輝ける統計学～

場所：１号館１５階　村上研究室

データの集積と多様化

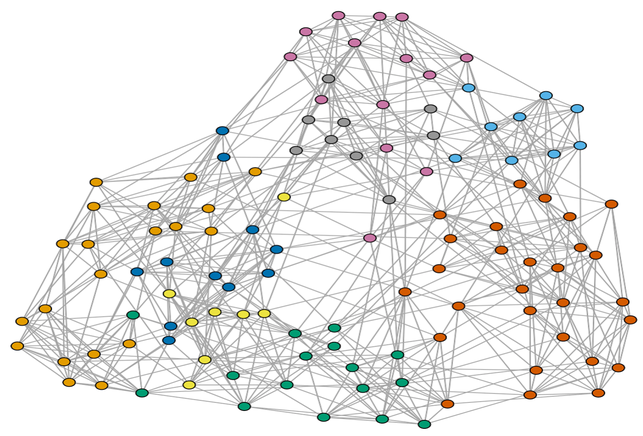
情報技術の著しい発展に伴い，莫大なデータの蓄積が可能となっている．例えば，

- 通信のログデータ (例：通信遅延時間データ)
- センシングデータ (例：環境データ)

などが代表的であり，容量，更新頻度，多様性などを主な特徴とするビッグデータとして注目を集めている．しかし，統計学で代表的な解析法をビッグデータへそのまま適用するとこれらの特徴が弊害となる可能性がある．また，得られるデータは数値だけでなく，

テキスト，画像，音声，グラフ

などがあるが，これらは直接的な統計処理が難しく，何らかの特徴量を求める必要がある．



グラフ形式データ

シンボリックデータ解析

データの規模が大きくなるにつれ，適度にデータを要約することが解析する上で重要となる．シンボリックデータ解析とは，従来のデータ構造の枠組みを一般化し，極力情報量を落とすことなく幾つかの集合や区間などに集約し，解析する方法である．

～～～区間値データの例～～～

区間値データとは，データの実現値を直接観測するのではなく，区間形式で得られるものである．例えば，ある気象観測センターによって1時間ごとの気象観測値

$$(\text{気温}(℃),\ \text{湿度}(\%),\ \text{風速}(m/s),\ \text{降水量}(mm/h))$$

が測定されるとする．各年の夏（6月～8月）のデータを集約すると

$$I_1 = ([19, 34],\ [60, 87],\ [1, 15],\ [0, 14]),$$
$$I_2 = ([23, 37],\ [64, 85],\ [2, 13],\ [2, 23]),$$
$$\vdots$$
$$I_n = ([18, 31],\ [53, 84],\ [1, 11],\ [1, 12])$$

として夏の気象の特徴を表現することが可能である．

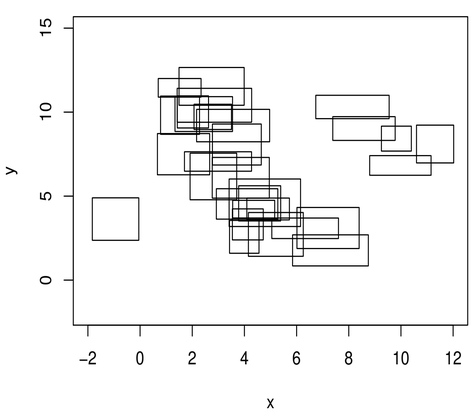
区間値データに対する密度推定

統計学において確率密度関数は重要であり，ノンパラメトリックな密度推定法は様々な分野で応用されている．区間値データ $\{I_1, I_2, \ldots, I_n\}$ に対する密度推定量を

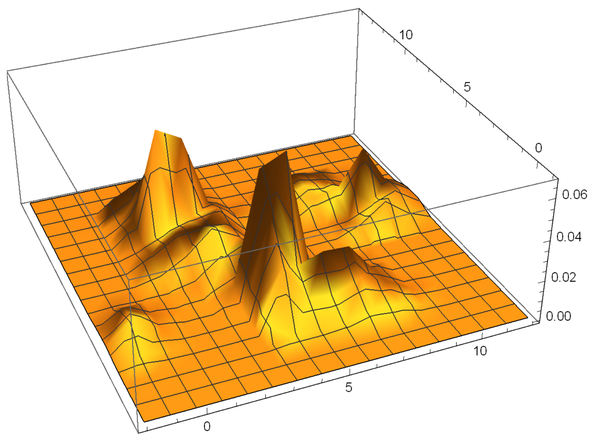
$$\hat{f}_n(\boldsymbol{x}) = \frac{1}{n}\sum_{k=1}^{n} |\hat{\Sigma}_k|^{-1/2} K\left(\hat{\Sigma}_k^{-1/2}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_k)\right)$$

と定義する．ただし，K は関数であり，$\hat{\boldsymbol{\mu}}_k$ と $\hat{\Sigma}_k$ は各区間値データの局所的な位置関係などを考慮して定まる値である．

～～～区間値データに対する密度推定の例～～～



区間値データ



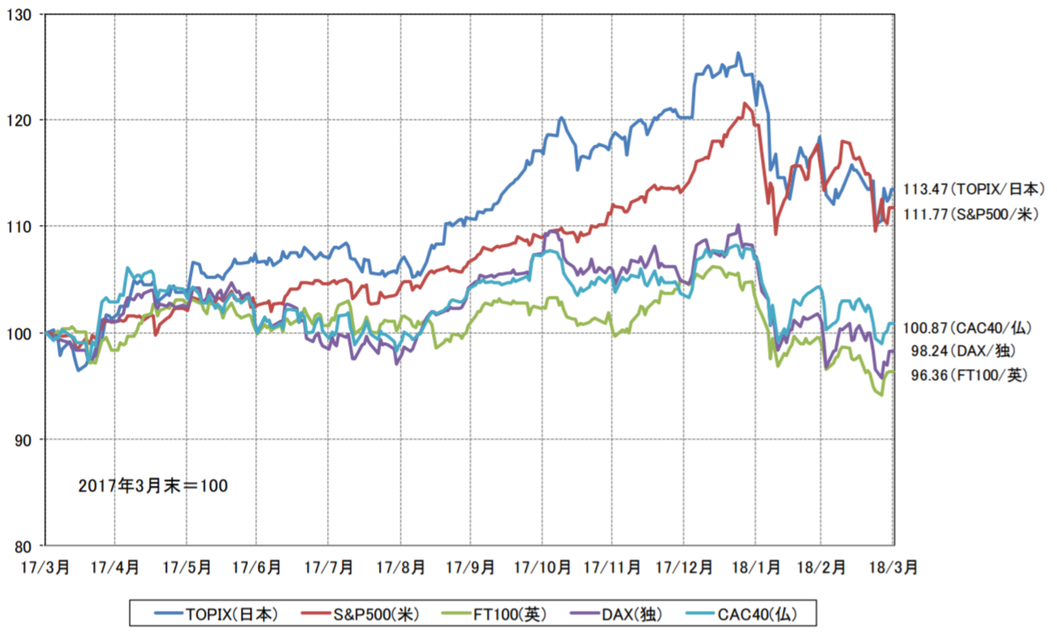
密度推定量

株価指数に対する分析

株価指数とは，株式の相場の状況を示すために個々の株価を一定の計算方法で総合し，数値化したものである．例えば，

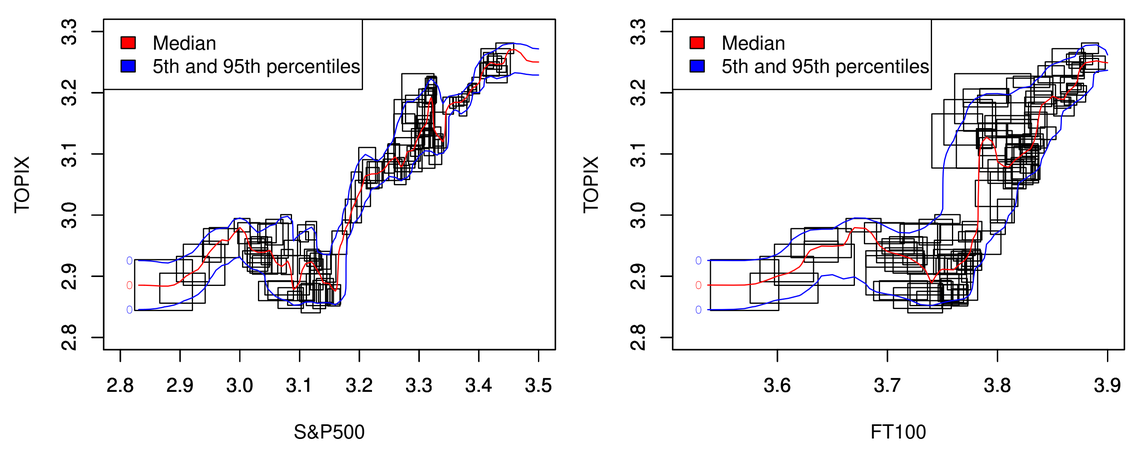
TOPIX（日本），S&P500（アメリカ），FT100（イギリス）

などがある．



株価指数の推移

下図は，リーマンショック以降の2009年1月から2018年6月までの各月の株価指数の [安値, 高値] を区間値として，株価指数間の関係を表したものであり，赤線と青線は密度推定によって推定されたパーセンタイルを示している．



※ 図は対数スケールで表示している．

～～～結果と考察～～～

株価指数について，

- 両方とも基本的に右肩上がりの関係になっている
- 信頼区間は日本とアメリカの方が狭く，日本はイギリスよりもアメリカの方が経済において強く依存している

という結果が得られた．

データを区間に集約して解析することによって今まで見えてなかったデータの構造や関係が見えてくる！