

3.4 ヒストリカル法

3.4.1 ヒストリカル法の概要と計算ステップ

- やりたいこと：マーケットの変動のVaR（下 $100\alpha\%$ 点）を求めたい。
 - そのためには**マーケット変動が従う確率分布**を求める必要がある。
 - ヒストリカル法とは、**過去のマーケット変動のデータを用いてノンパラメトリックに確率分布を作成する手法**。
（ヒストグラムを作るだけですけど。）
- メリット：
 - マーケット変動の従う確率分布にパラメトリックな仮定を入れる必要がない（正規分布や対数正規分布など）
 - 単純（そりゃ中学で習うヒストグラムですから。）
- デメリット：
 - 過去のデータをどのくらい信頼しても良いのか分からない。
（50年前の変動と現在の変動は絶対違うし、というかそもそも**マーケットの時系列は非定常**）
 - データが少ない場合、外れ値にかなり敏感

導出方法

- 得られたマーケットのデータ（時系列データ） $\{\xi_t\}_{t=1}^T$ を時刻 t に独立な系列と考え、その値の大きい順に並び替える。
 $\rightarrow \{\dot{\xi}_k\}_{k=1}^T$
- この系列の小さい方から順に $100\alpha\%$ のデータ点 $\dot{\xi}_{T\alpha}$ を $VaR_{100(1-\alpha)}$ とする。

$$\dot{\xi}_{T\alpha} = VaR_{100(1-\alpha)}$$

- 要はただの分位点？

3.4.2 ヒストリカル法に対する批判と反論

- 批判1：過去のデータをそのまま使ったら、過去の変動の傾向が将来ずっと続くことになっちゃうでしょー
 - 反論：それはどの手法でもそうやる
- 批判2：データ量少ない時に精度が低いのはなー

- 反論：それだって他の手法も同じで、他の手法より明確に精度が低い根拠はないぞ。（まあ実際にやってみると精度低いけど...）
- 批判3：データ数が霧のいい数字じゃないと、分位点を出せないじゃないかー
 - 反論：でもデータ数がそこそこあればそれは大きな問題じゃないし、むしろデータ数が多い時はパラメトリックな手法より精度がいい研究もあるヨ

それでは、なぜ金融機関で一般的に使われていないのか？

→ 理論が簡単すぎてつまんないから。

（著者、ヒストリカル法絶対大好きでしょ...）

3.4.4 ブートストラップ法

ブートストラップ法とは...

ヒストリカル法において作成されたノンパラメトリックな確率分布（経験分布とも）から再サンプリングし、新しくデータセットを作成する、統計的シミュレーションの一種

1. 最初に得られたデータセット $\{x_i\}_{i=1}^N$ から経験分布 \mathbb{P}_x を作成する。
2. \mathbb{P}_x から N 回サンプリングし、得られた新しいデータセット $\{x_i^{(1)}\}_{i=1}^N$ を活用する。
3. 2. を繰り返し行うことで、一つのデータセットからたくさんの（擬似的な）データセット $\{x_i^{(1)}\}_{i=1}^N, \{x_i^{(2)}\}_{i=1}^N, \dots, \{x_i^{(b)}\}_{i=1}^N$ を獲得できる。

なんでこんなことするのか？

一言で言えば、「**統計量の分布が擬似的に観測できる**」。

例えば得られたデータセット $\{x_i\}_{i=1}^N$ から、背後にある母集団の期待値を導出する統計量 $\hat{\theta}$ は、

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^N x_i$$

と出せるが、この統計量の分布は一般的には分からない。（もちろん正規分布みたいな簡単な分布なら出せるけど。）

でも、ブートストラップ標本が b セットあれば、統計量 $\hat{\theta}$ も b 個得られる。つまりこの b 個の統計量のデータセット $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \dots, \hat{\theta}^{(b)}$ から、経験分布を作成することで、統計量 $\hat{\theta}$ の分布を導出できる。

期待値だとあんまり旨味がある話に聞こえないけど、例えばVaR、ランダムフォレストの分割基準、異常検知における検出基準など、その確率分布の導出がクソだるい統計量の分布が楽に導出できてしまう。

ブートが有効な場面：

- 統計量が非線形で母集団の従う分布も非対称
- 統計量の導出がそもそも大変（学習によって得られた統計量など）
- データ数が少ない

統計量の分布がわかると、AICや信頼係数などが出せて嬉しい。

デメリットは最初のデータセット $\{x_i\}_{i=1}^N$ の持つバイアスに割と敏感なこと。研究で使うなら、人工データでブートストラップ統計量へのバイアスの影響を調べておくのが吉。

あと著者曰く、VaRの導出においては経験的にブートストラップは向かないとのこと。（じゃあなんで紹介したんだ）