角度データのための扇形ヒストグラムの漸近的性質

和歌山県データ利活用推進センター研究員　　鶴田靖人
金沢大学経済学経営学系　　寒河江雅彦

角度データ (風向等) は周期性をもつために実数直線上のデータと位相構造が異なり，単位円周上の点として表される．そのために角度データを扱うための独自の統計学が発展し、最近では方向統計学と呼ばれる統計学の 1 つの分野となっている．方向統計学では角度データ (変数)$\Theta \sim f(\theta)$ を周期性を持つ密度関数 $f(\theta)$ ($f(\theta) = f(\theta + 2\pi)$) に従うと定義する．

本稿では角度データのためのヒストグラムを密度推定量と考え，その理論的性質を議論する．角度データのヒストグラム推定で最も用いられているのはローズダイアグラムである．ローズダイアグラムは，データが入る区間であるビン $B_k := [t_k, t_{k+1}) \in [-\pi, \pi)$ を中心角として持つ面積 v_k/n (B_k の相対頻度) の扇形 S_k を原点周りに並べることで定義される．一般的に各ビン B_k の中心角の大きさはすべて h（等角度）とする．このとき，扇形の面積の公式から S_k の半径は $r_k = \sqrt{2v_k/(nh)}$ である．

分布関数 $F(\theta)$ を原点 O から伸びる線分 $r_f(\theta)$ の通過領域からなる扇形として与える．$f(\theta) = dF(\theta)/d\theta$ なので扇形の面積の公式より $r_f(\theta) = \sqrt{2f(\theta)}$ となる．ここで，ローズダイアグラムの半径 r_k を $r_f(\theta)$ の推定量と考え，ローズダイアグラム推定量を

$$\hat{r}(\theta; h) := \sqrt{2v_k/(nh)}, \quad \theta \in B_k$$

と定義する．通常のヒストグラム推定量 $\hat{f}(\theta; h) := v_k/(nh)$ を用いると $\hat{r}(\theta; h) = \{2\hat{f}(\theta; h)\}^{1/2}$ となる．ローズダイアグラム推定量の誤差基準として，平均積分二乗誤差 (MISE) $\mathrm{MISE}[\hat{r}(\theta; h)] := \mathrm{E}\left[\int_{-\pi}^{\pi} \{\hat{r}(\theta; h) - r_f(\theta)\}^2 d\theta\right]$ を採用する．ちなみに，$\mathrm{MISE}[\hat{r}(\theta; h)]$ は，$\hat{r}(\theta; h)$ の定義から Hellinger 距離 $\mathrm{HD}[\hat{f}(\theta; h)] := \int_{-\pi}^{\pi} \{\hat{f}(\theta; h) - f(\theta)\}^2 d\theta$ の期待値に対応することが容易に分かる．本稿の主要な結果は次の 3 点である：

- $\mathrm{MISE}[\hat{r}(\theta; h)]$ の導出 ($\mathrm{MISE}[\hat{r}(\theta; h)] = O(n^{-2/3})$).
- MISE 基準に基づくビン幅 h の推定量の提案.
- frequency polygon を応用したローズダイアグラムの改良 ($\mathrm{MISE} = O(n^{-4/5})$).

当日の発表ではこれらの結果と合わせて，$\mathrm{MISE}[\hat{r}(\theta; h)]$ と Hellinger 距離の対応関係について理論的に考察する．また，ローズダイアグラム推定量に関する数値実験の結果は当日発表する．

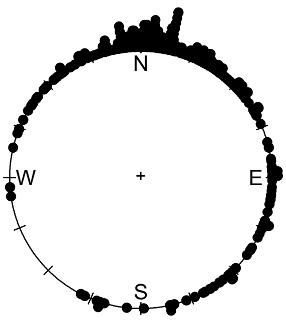

図 1　風向を表す Wind データ (n=310)．Wind データは，統計ソフト R の `circular` パッケージから取得できる．

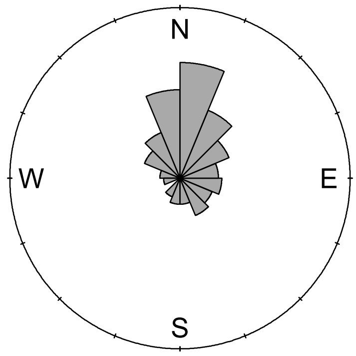

図 2　Wind データのローズダイアグラム．