

角度データの統計処理基礎

2012/1/7 第4回 定量生物学の会 チュートリアル

石原秀至^{1,2}

¹東京大学大学院総合文化研究科

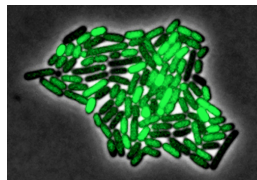
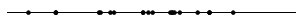
²JSTさきがけ

※お借りしたデータなどの都合で一部変更しています

リニアデータと角度データ

リニアデータ

$$\{x_1, x_2, \dots, x_n\}$$

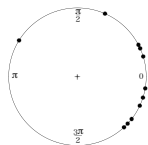


GFPのシグナル強度

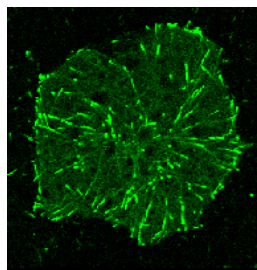
Credit: Elowitz lab

角度データ

$$\{\theta_1, \theta_2, \theta_3, \dots, \theta_n\}$$



$$\theta + 2\pi = \theta$$



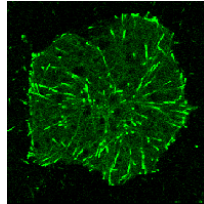
微小管の(+)端の移動方向
EB3-GFPコメットの移動方向

Shindo et al., PLoS one, 2008

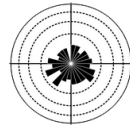
movie

生物学における角度データの例

微小管の(+)端の移動方向

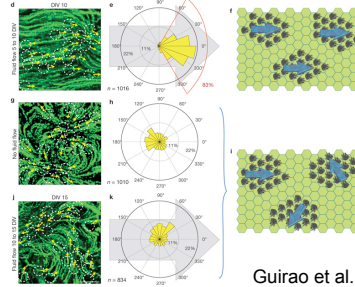


コメットの進行方向の角度分布



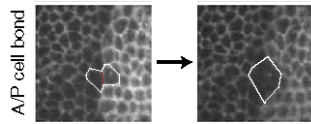
Shindo et al., PLoS one, 2008

繊毛のbeatingによる流れの方向

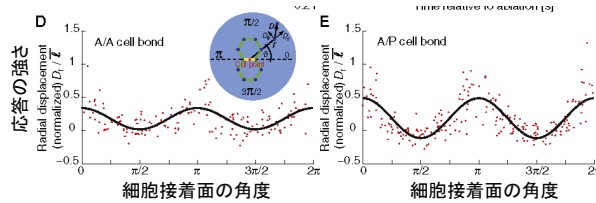


Guirao et al., NCB, 2010

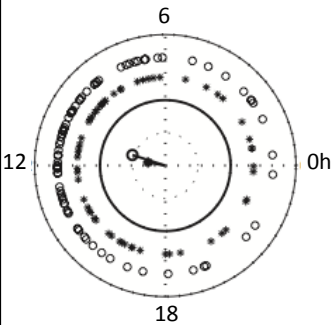
細胞接着面の角度とレーザー切断後の頂点の移動速度(張力)の関係



Landsberg et al., Current Biol, 2009



位相 = 角度



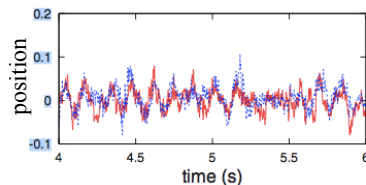
ショウジョウバエ個体の活動ピーク時間

J. D. Levine et al. Science (2002)

* isolated (平均(後述))
 ○ in group

→ 2群比較(検定)
 二つの条件化でのデータは同じ傾向を示すと言えるか?

振動的な時系列データ → 位相振動子としての解析



眼球運動 Romano et al. Chaos 2010

左眼と右眼の運動が同期しているといえるか?

二つの時系列のシンクロ同定, 位相応答曲線推定, etc..

チュートリアル: 角度データの統計処理基礎

・生物学における角度データ

形態の特徴付けなど、いろいろなところでは会う。
わりと混乱する(した、しているを見た)。一度整理しておくと楽/便利。

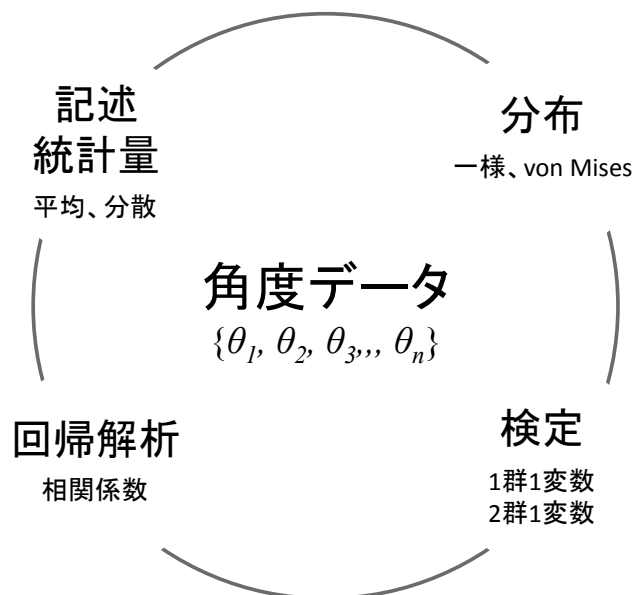
・角度統計

昔からある
あまり知られていない?(日本語の文献は少ない)
神経系ではちらほら

・方針

ユーザーの立場からプラクティカルに説明する
基本的な考え方だけを説明
キーワードをメモして、詳細は文献を参照してください。(厳密でないです)
2次元の場合だけ
こみいった突っ込みには答えられませんのであしからず。(助けてください)

マップ



角度データ統計 (circular/directional statistics)

統計量(平均、分散など)、分布、検定

例1: 繊毛が生えている向き

N個の角度データが得られたとき、その平均は？分散は？ **記述統計量**

N個の角度データが得られたとき、ある方向に偏っていると言えるのか？ **検定**

例2: 位相

ある時刻になるとおなかがへりやすいと言えるのだろうか？ **検定**

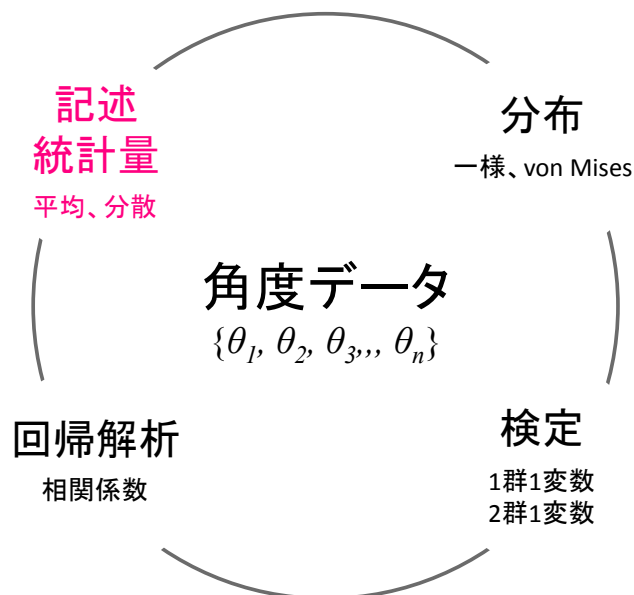
例3: 細胞分裂の向き

$\theta = \theta + 180^\circ$ (軸性)のとき、平均と分散は？ **記述統計量**

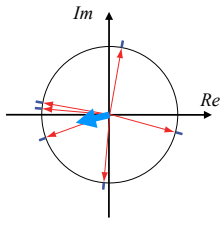
例4: 細胞接着面の角度と張力に相関はあるだろうか？

データ (T_i, θ_i) があったとき、 θ - T 間の関係 **相関/回帰分析**

マップ



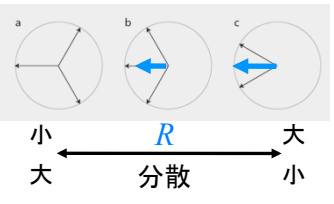
分散・標準偏差



平均
↓
角度 θ 、長さ R のベクトル

?

$$Re^{i\theta} = \frac{1}{N} \sum_j e^{i\theta_j} = \langle e^{i\theta_j} \rangle$$



小 ← R → 大
分散

分散
(circular variance)

$$V \equiv 1 - R \quad (0 \leq V \leq 1)$$


標準偏差
(standard deviation)

$$S \equiv \sqrt{-2 \log(R)}$$

角度データのばらつきが大きいと R が小さく、
ばらつきが小さいと R は大きい

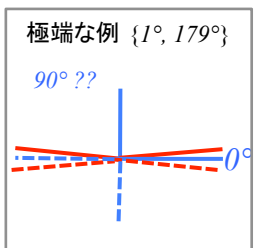
軸性角度: θ と $\theta+180^\circ$ が見分けられない場合

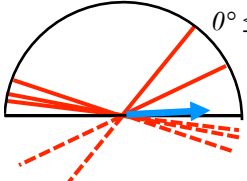
例: 細胞分裂の方向



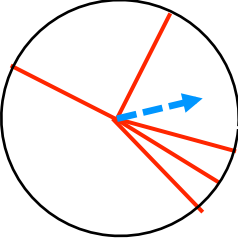
$\{170^\circ, 175^\circ, 160^\circ, 65^\circ, 35^\circ\}$
 $0^\circ \leq \theta < 180^\circ$

極端な例 $\{1^\circ, 179^\circ\}$
 $90^\circ??$





$\times 2$



平均は?

✗ $Re^{i\theta} = \langle e^{i\theta_j} \rangle$

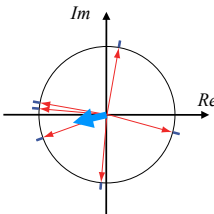
○ $Re^{i2\theta} = \langle e^{i2\theta_j} \rangle$

いったん角度を2倍して平均値を求め、それを2で割る

$\{170^\circ, 175^\circ, 160^\circ, 65^\circ, 35^\circ\}$
↓
 $\{340^\circ, 350^\circ, 320^\circ, 130^\circ, 70^\circ\}$

平均値: $\theta = 6.5^\circ$
↑
平均値: $2\theta = 12.9^\circ$

ここまでのまとめ



$(R \cos \Theta, R \sin \Theta) = \frac{1}{N} \left(\sum_j \cos \theta_j, \sum_j \sin \theta_j \right)$
角度 Θ
長さ R のベクトル

$R e^{i\Theta} = \frac{1}{N} \sum_j e^{i\theta_j} = \langle e^{i\theta_j} \rangle$
複素平面における表記

平均

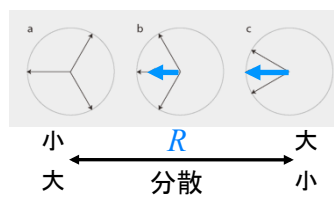
Θ

分散

$V \equiv 1 - R \quad (0 \leq V \leq 1)$

標準偏差

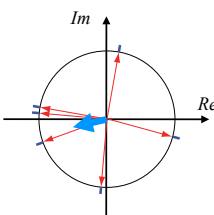
$S \equiv \sqrt{-2 \log(R)}$



小 \xleftarrow{R} 大

大 $\xrightarrow{\text{分散}}$ 小

ここまでのまとめ



$(R \cos \Theta, R \sin \Theta) = \frac{1}{N} \left(\sum_j \cos \theta_j, \sum_j \sin \theta_j \right)$
角度 Θ
長さ R のベクトル

$R e^{i\Theta} = \frac{1}{N} \sum_j e^{i\theta_j} = \langle e^{i\theta_j} \rangle$
複素平面における表記

平均

Θ

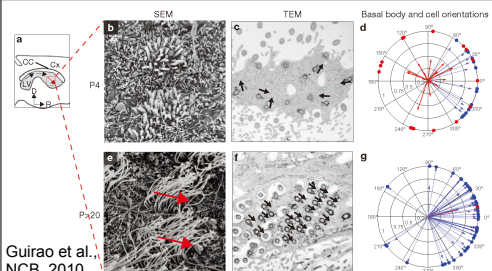
分散

$V \equiv 1 - R \quad (0 \leq V \leq 1)$

標準偏差

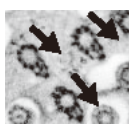
$S \equiv \sqrt{-2 \log(R)}$

SEM TEM Basal body and cell orientations



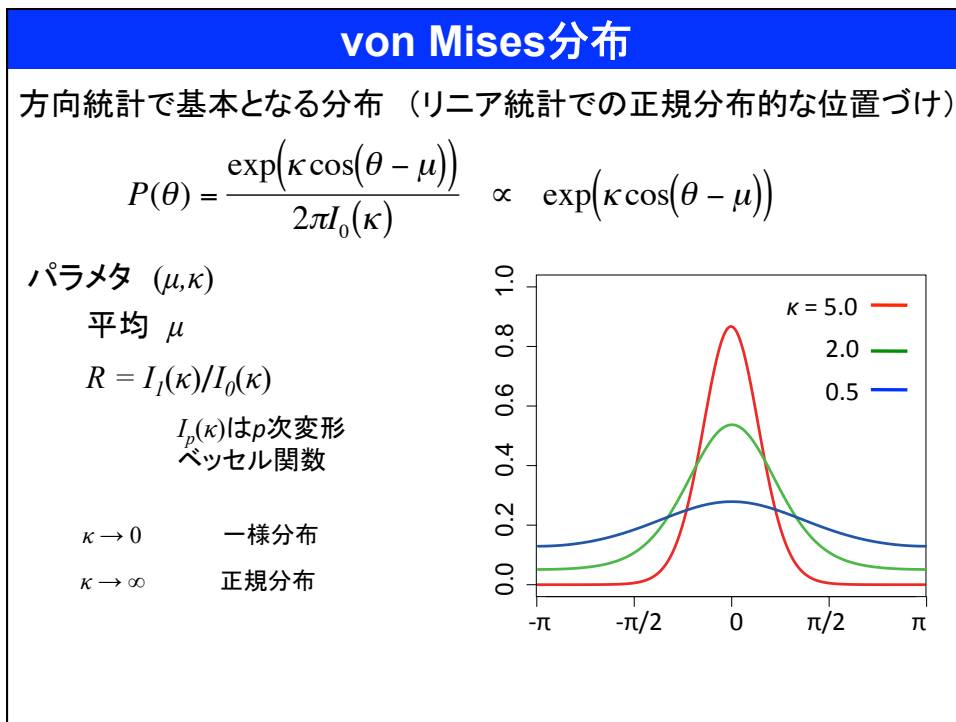
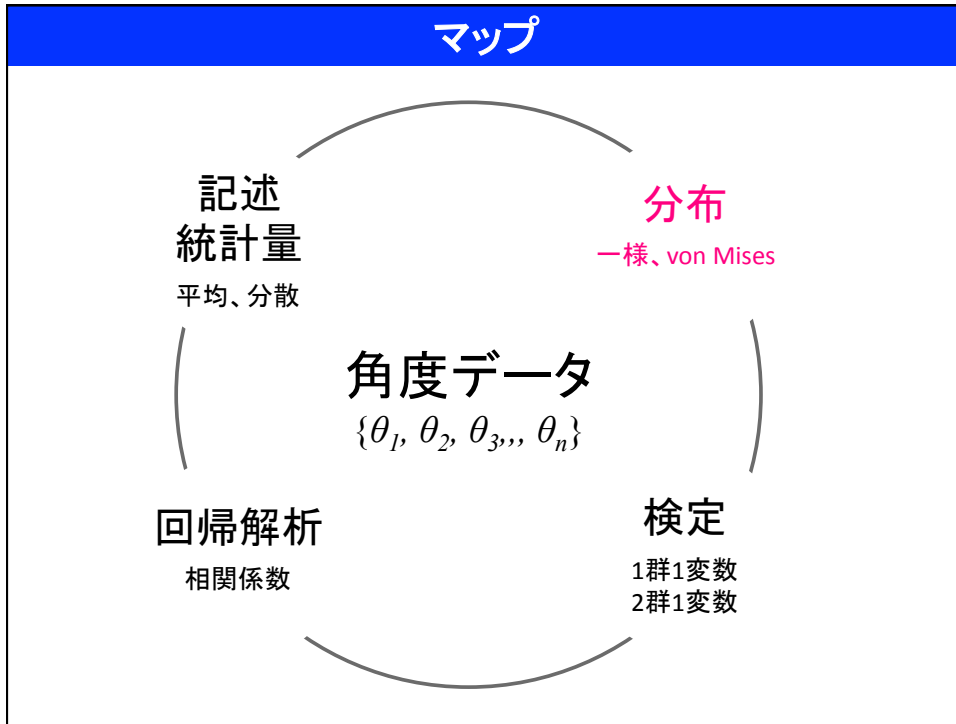
Guirao et al., NCB, 2010

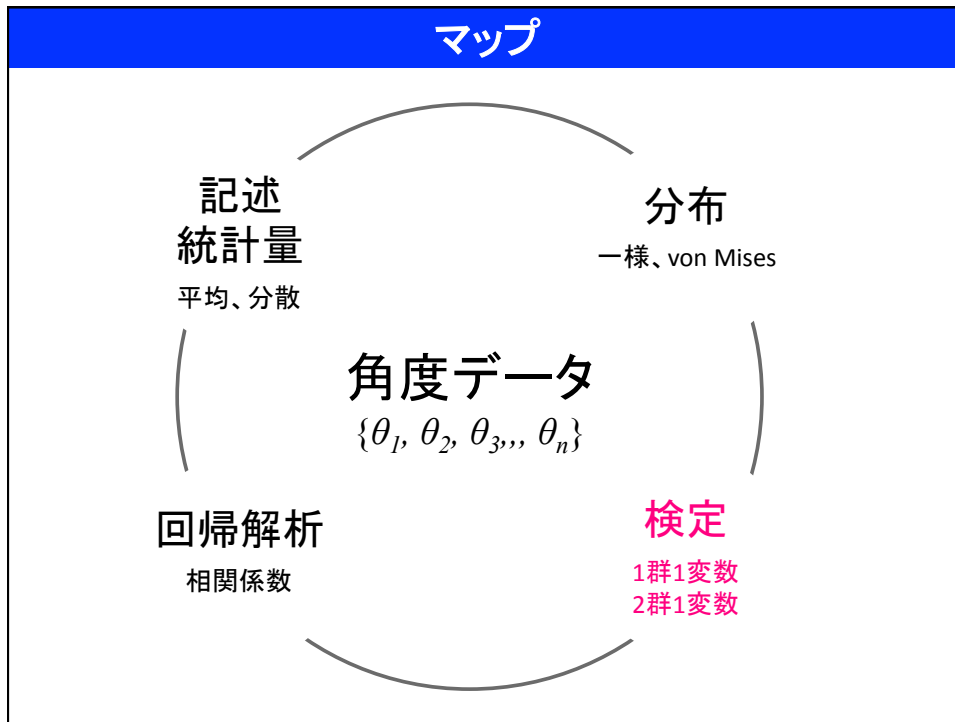
たくさんの繊毛/1細胞
各繊毛のBasal bodyの角度を測定
細胞ごとに角度 θ 長さ R のベクトルを表示
P4(生まれて4日後)とP20で比較



$R_{P4} < R_{P20}$

$V_{P4} > V_{P20}$





検定の手続き(例: 2標本検定)

問い: 焼きじゃがいもに味噌をつけて食べると早死にするのか?

データ:

焼きじゃがいもに味噌をつけて食べた人の死亡年齢 N_A , 平均(E_A), 分散(V_A)

焼きじゃがいもに味噌をつけて食べなかった人の死亡年齢 N_B , 平均(E_B), 分散(V_B)



検定統計量を計算



帰無仮説(同じ分布に従う)のもとで検定統計量の出現確率 p を計算。
有意水準(たとえば、 $p < 0.01$)で帰無仮説を棄却できるか否かを判定

角度データの代表的な検定

Rayleigh test
 角度データには偏りがあるか？
 ある角度に偏っているのか？

Kuiper test
 角度データはvon Mises分布に従っているのか？

Mardia-Watson-Wheeler test
 2群のデータは同じ分布に従っているのか？

Rayleigh test: 角度データの異方性

A. 角度データに偏り(異方性)があるといえるか？

① R を計算 $Re^{i\theta} = \langle e^{i\theta_i} \rangle$ 角度 θ 、長さ R のベクトル

② R が大きければ一様分布から外れていると言える。

一様分布(帰無仮説)のもとでは、サンプル数 n の時に $Z = nR^2$ が出る確率は

$$P = e^{-Z} \left(1 + \frac{2Z - Z^2}{4n} - \frac{24Z - 132Z^2 + 76Z^3 - 9Z^4}{288n^2} \right) \sim e^{-Z} \quad (\text{p値}) \quad \text{なので、}$$

Z が大きければ「異方性がある」と主張できる (帰無仮説を棄却できる)

SEM: b (P4), e (P20)
 TEM: c (P4), f (P20)
 Basal body and cell orientations: d (P4), g (P20)

たくさんの繊毛/1細胞
 各繊毛のBasal bodyの角度を測定
 細胞ごとに角度 θ 長さ R のベクトルを表示
 P4(生まれて4日後)とP20で比較

$R_{P4} < R_{P20}$

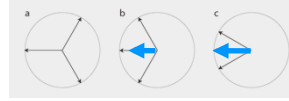
$P \geq 0.05$ $P < 0.05$

Guirao et al., NCB, 2010

Rayleigh test: 角度データの異方性

A. 角度データに偏り(異方性)があるといえるか？

① R を計算 $Re^{i\theta} = \langle e^{i\theta_i} \rangle$ 角度 θ 、長さ R のベクトル



② R が大きければ一様分布から外れていると言える。

一様分布(帰無仮説)のもとでは、サンプル数 n の時に $Z = nR^2$ が出る確率は

$$P = e^{-Z} \left(1 + \frac{2Z - Z^2}{4n} - \frac{24Z - 132Z^2 + 76Z^3 - 9Z^4}{288n^2} \right) \sim e^{-Z} \quad (\text{p値}) \quad \text{なので、}$$

Z が大きければ「異方性がある」と主張できる (帰無仮説を棄却できる)

B. ある角度 θ_0 に偏っているといえるか？ (角度 θ_0 を指定、V-test)

① $R_\theta = R \cos(\theta - \theta_0)$ を計算

② R_θ が大きければ角度 θ_0 に偏っている度合いが大きいといえる。

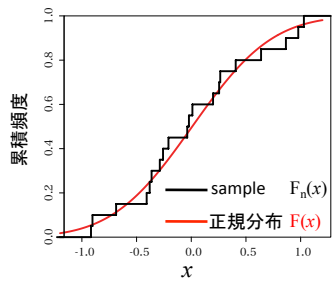
一様分布(帰無仮説)のもとでの $Z = (2n)^{1/2} R$ が出る確率 (p値) をもとに
帰無仮説を棄却できるか否かを判定する

Kuiper test: データが von Mises 分布に従っているか

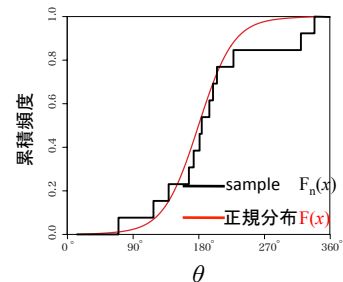
Kolmogorov-Smirnov (KS) 検定の角度データ版
 1標本の適合度検定 (ある分布に従っているのか?) 従っていないと言えるか?
 (2標本が同じ分布からサンプルされているのか?) 異なっていると言えるか?

KS検定 例 $\{x_1, x_2, x_3, \dots, x_n\}$
 (リニアデータ) は正規分布に従っているか?

Kuiper検定 $\{\theta_1, \theta_2, \theta_3, \dots, \theta_n\}$
 (角度データ) は von Mises 分布に従っているか?



検定量 $D = \max |F_n(x) - F(x)|$



検定量 $V = \max (F_n(x) - F(x)) + \max (F(x) - F_n(x))$

帰無仮説(二つの分布は一致)のもとで、 V の出る確率(p値)を評価

Mardia-Watson-Wheeler test: 2群は同じ分布に従っているか

U検定 (リアデータ) $\{x_1, x_2, x_3, \dots, x_n\}$ と $\{y_1, y_2, y_3, \dots, y_m\}$ \rightarrow $n+m$ 個のデータを混ぜて **順位付け**
 n 個 < m 個

A群 $n=6$ 2 3 6 7 10 14 $R = 42$ (A群の**順位和**)
 B群 $m=8$ 1 4 5 8 9 11 12 13 $U = 27$

$\{x\}, \{y\}$ が同じ分布に従うとならば、実現された順位づけが出る確率(**p値**)が計算できる。
p値が小さければ、帰無仮説「同じ分布から得られた」を棄却できる。

(標本数が多い場合には) 検定量 $U = nm + n(m+1)/2 - R$ (R は x の順位総和)から判断できる。

Mardia-Watson-Wheeler test はMann-WhitneyのU検定の角度データ版

Mardia-Watson-Wheeler test: 2群は同じ分布に従っているか

MWW検定 (角度データ) $\{\theta_1, \theta_2, \theta_3, \dots, \theta_n\}$ と $\{\psi_1, \psi_2, \psi_3, \dots, \psi_m\}$
 n 個 < m 個

\rightarrow $n+m$ 個のデータを混ぜて小さい順に $0-2\pi$ で等間隔に並べる
 \rightarrow $\left\{ \frac{2\pi\theta_1}{n+m}, \frac{2\pi\theta_2}{n+m}, \dots, \frac{2\pi\theta_n}{n+m} \right\}$ $\left\{ \frac{2\pi\psi_1}{n+m}, \frac{2\pi\psi_2}{n+m}, \dots, \frac{2\pi\psi_m}{n+m} \right\}$ θ_j, ψ_j は0から $n+m-1$ の整数

● A群 ($n=6$)
● B群 ($m=8$)

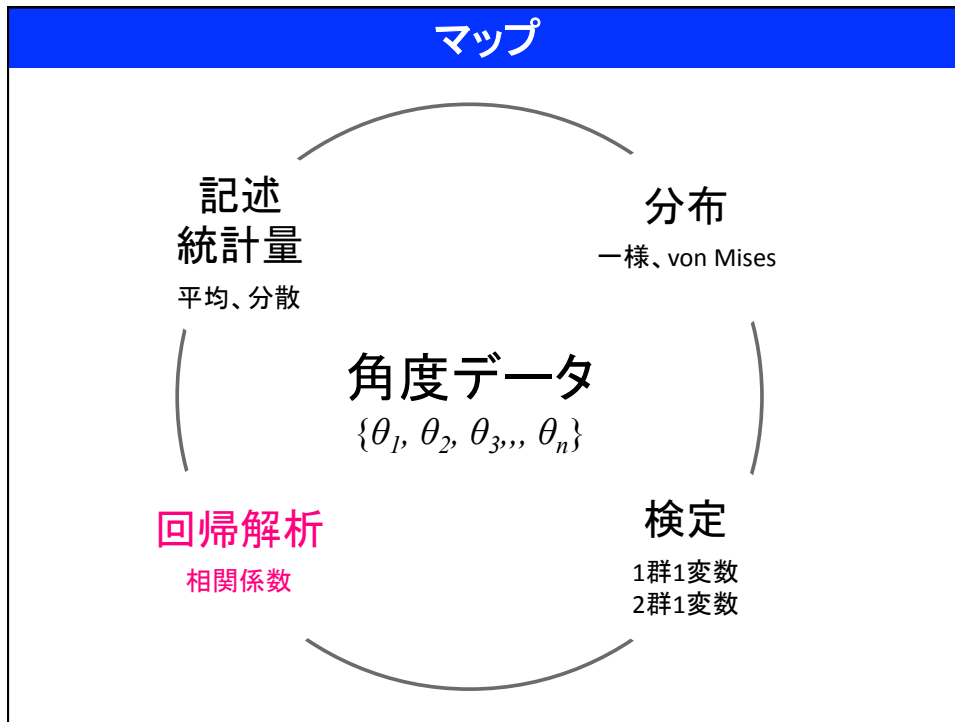
等間隔に
並べ直す

R小

R大

従っていない時は偏る

A群に関して**R**を計算 標本数が小さい時は直接確率を計算する
Rの大きさに基づいて帰無仮説(同じ分布から得られた)を棄却するかどうかを判定する



相関・回帰分析

Circular-circular correlation
例: 細胞の向きと分裂方向に相関はあるのか?

Linear-circular correlation
例: 細胞接着面の角度とレーザー切断に対する応答の強さ(張力)に相関があるのか?

一次フィッティング

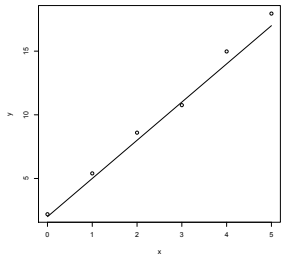
最小二乗法 $L_{ll}(a,b), L_{cl}(a,b,c)$ を最小にする (a,b,c) を求める.

リニア-リニアデータ (x_j, y_j)

$$y = a + bx$$

$$L_{ll}(a,b) = \sum_j |y_j - a - bx_j|^2$$

b が大きければ x 依存性が大きい



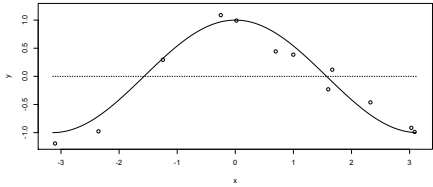
リニア-角度データ (S_j, θ_j)

$$S = a + b' \cos(\theta - \mu)$$

$$= a + b \cos \theta + c \sin \theta$$

$$L_{cl}(a,b,c) = \sum_j |S_j - a - b \cos \theta_j - c \sin \theta_j|^2$$

b' が大きければ角度依存性が大きい



一次フィッティング

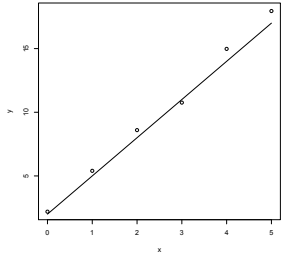
最小二乗法 $L_{ll}(a,b), L_{cl}(a,b,c)$ を最小にする (a,b,c) を求める.

リニア-リニアデータ (x_j, y_j)

$$y = a + bx$$

$$L_{ll}(a,b) = \sum_j |y_j - a - bx_j|^2$$

b が大きければ x 依存性が大きい



リニア-角度データ (S_j, θ_j)

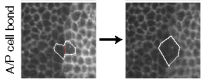
$$S = a + b' \cos(\theta - \mu)$$

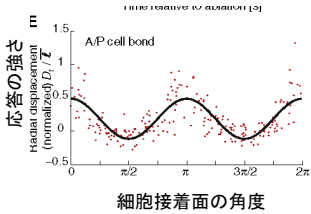
$$= a + b \cos \theta + c \sin \theta$$

$$L_{cl}(a,b,c) = \sum_j |S_j - a - b \cos \theta_j - c \sin \theta_j|^2$$

b' が大きければ角度依存性が大きい

A/P cell bond





Landsberg et al., Current Biol, 2009

相関係数

■ リニア-リニアデータ対 (x_j, y_j) についての Pearson 相関係数

$$r_{xy} = \frac{\langle \Delta x_j \Delta y_j \rangle}{\sqrt{\langle \Delta x_j^2 \rangle} \sqrt{\langle \Delta y_j^2 \rangle}}$$

■ 角度-リニアデータ 対 (θ_j, S_j) に対して1次のfitting

$$\begin{aligned} S &= a + b' \cos(\theta - \mu) \\ &= a + b \cos \theta + c \sin \theta \quad (\diamond) \end{aligned}$$

式(\diamond)を変数 $(\cos \theta, \sin \theta)$ に対する2変数線形fittingだと思つと、交互作用を考慮した相関係数を考えればよい。

$$\rho_{\theta, S} = \sqrt{\frac{r_{cS}^2 + r_{sS}^2 - 2r_{cS}r_{sS}r_{cs}}{1 - r_{cs}^2}}$$

r_{cS} : $\cos\theta$ - S のPearson相関

r_{sS} : $\sin\theta$ - S のPearson相関

r_{cs} : $\cos\theta$ - $\sin\theta$ のPearson相関

r_{cs} : $\cos\theta$ - $\sin\theta$ のPearson相関

$\rho_{\theta, S}$ がどれくらい大きければ相関を主張できるのか？

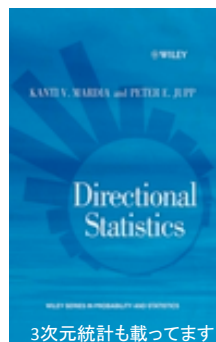
→ *Bootstrap*(詳細は文献参照)

参考文献 (年会web pageに掲載されています)

Fisher
**Statistical Analysis
of Circular Data**



Mardia & Jupp
Directional statistics



Batschelet
**Circular statistics
in biology**

絶版

日本語の文献は少ない

「逆」引き統計学実践統計テスト100(カンジ著、池谷・久我訳)に検定がいくつか載っています

実装

MATLAB

circular statistics toolbox

by Philippe Berens

Circular Statistics Toolbox (Directional Statistics)

by Philipp Berens
08 Apr 2006 (Updated 19 Apr 2011)

Compute descriptive and inferential statistics for circular or directional data.

Editor's Notes:
This file was selected as **MATLAB Central Pick of the Week**

[Watch this File](#)

R

circular statistics package

Circular Statistics

※ 提供されている関数を見ると、角度統計で何が出来るかの参考になります。

