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　[2]JSTさきがけ

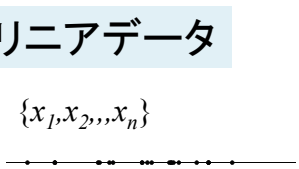
(お手伝い: 杉村　薫 京都大学iCeMS)

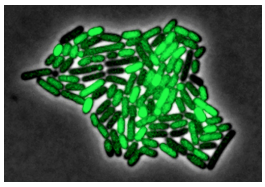
131208 ver.02 upload

リニアデータと角度データ

リニアデータ

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



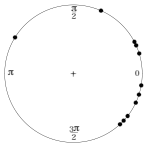


GFPのシグナル強度

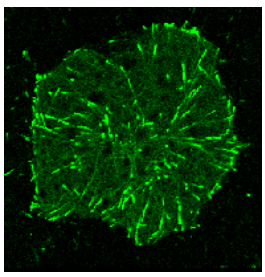
Credit: Elowitz lab

角度データ

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



$\theta + 2\pi = \theta$



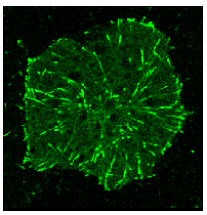
movie

微小管の(+)端の移動方向
EB3-GFPコメットの移動方向

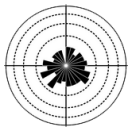
Shindo et al., PLoS one, 2008

生物学における角度データの例

微小管の(+)端の移動方向

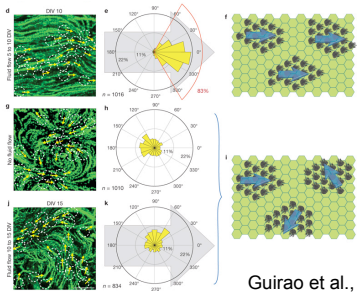


コメットの進行方向
の角度分布



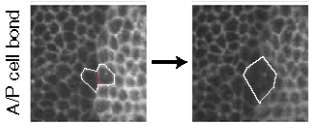
Shindo et al., PLoS one, 2008

繊毛のbeatingによる流れの方向

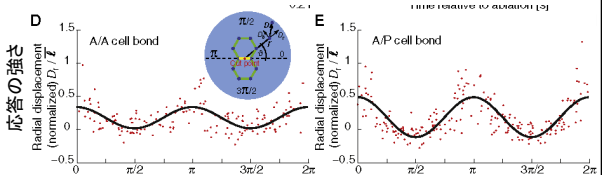


Guirao et al., NCB, 2010

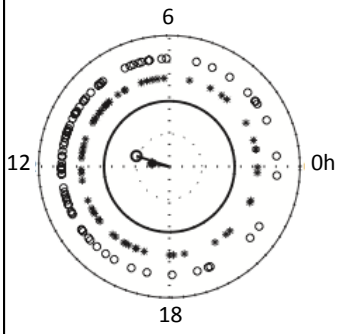
細胞接着面の角度とレーザー切断後の頂点の移動速度(張力)の関係



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

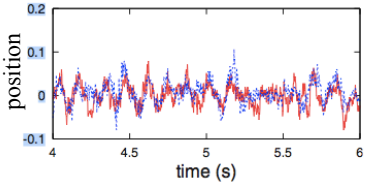
J. D. Levine et al. Science (2002)

＊ isolated
○ in group

(＊—— / ○—— 平均 (後述))

→ 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出会う。
わりと混乱する(した、しているのを見た)。一度整理しておくと楽/便利。

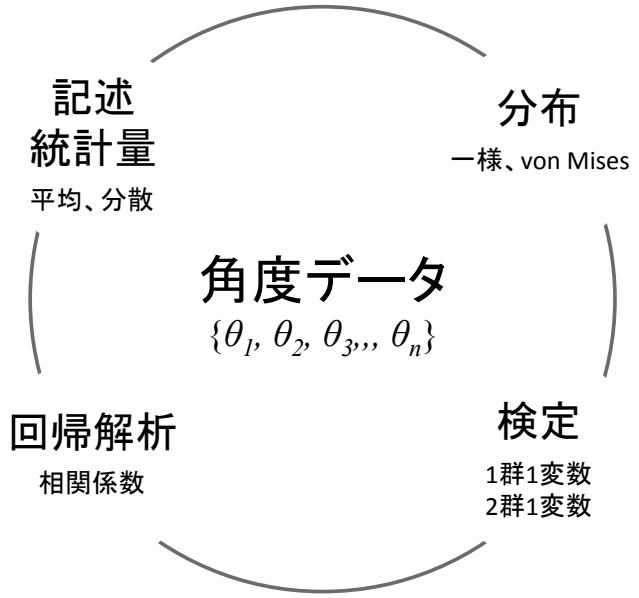
・角度統計

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**

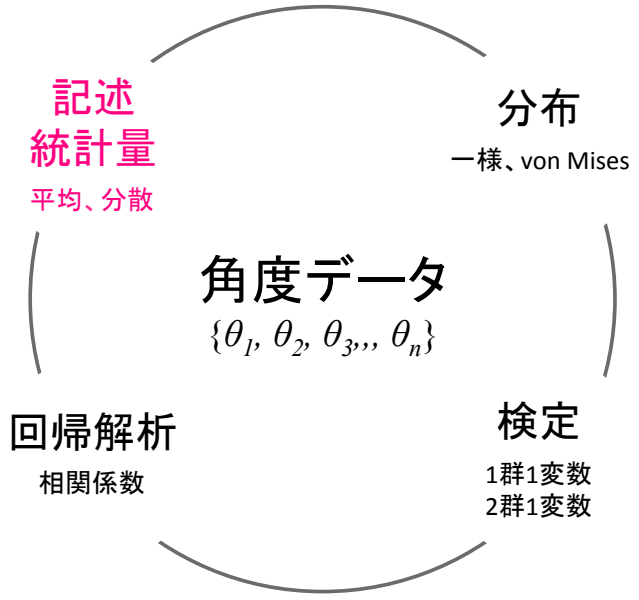
例3: 細胞分裂の向き

$\theta = \theta + 180°$ (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？

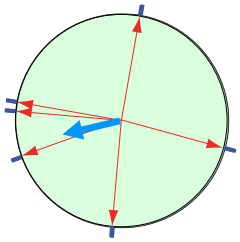
データ (T_i, θ_i) があったとき、θ-T 間の関係 **相関/回帰分析**

マップ

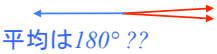


平均

例: 角度データ {80°, 170°,175°, 200°,265°, 345°}



極端な例 {1°, 359°}

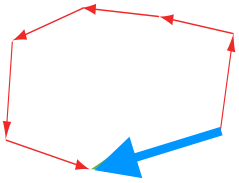


平均は180° ??

平均は0°

✖ (80°+170°+175°+200°+265°+345°)/6=206° ?

◯ ベクトルの平均をとる



$$(R\cos\Theta, R\sin\Theta) = \frac{1}{N}\left(\sum_j \cos\theta_j, \sum_j \sin\theta_j\right)$$

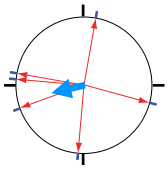
(× 1/N)

角度 Θ、長さRのベクトル

平均値: Θ=191°

記法の注①②

① 平均値は⟨•⟩で表す $\langle x_j \rangle \equiv \frac{1}{N}\sum_j x_j$



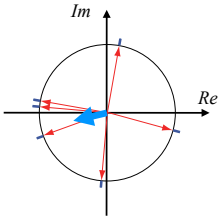
角度 Θ、長さRのベクトル

ベクトルの平均

$$(R\cos\Theta, R\sin\Theta) = \frac{1}{N}\left(\sum_j \cos\theta_j, \sum_j \sin\theta_j\right) = \left(\langle\cos\theta_j\rangle, \langle\sin\theta_j\rangle\right)$$

x成分 y成分　　x成分 y成分

② 複素平面だとおもうと便利



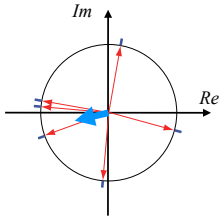
オイラーの公式

$e^{i\theta} = \cos\theta + i\sin\theta$　実数部がx成分、虚数部がy成分

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

同じ

分散・標準偏差

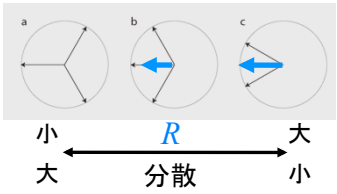


平均 ?

↓ ↓

角度 Θ、長さRのベクトル

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



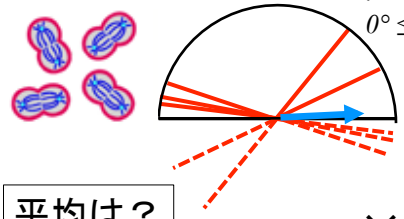
分散 (circular variance) $V \equiv 1 - R \quad (0 \le V \le 1)$

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 θ+180° が見分けられない場合

例: 細胞分裂の方向

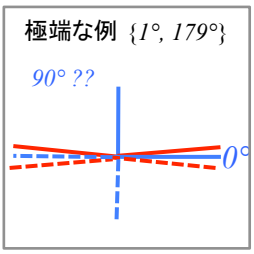


{ 170°,175°, 160°,65°, 35°}

$0° \le \theta < 180°$

$0° = 180°$

極端な例 {1°, 179°}

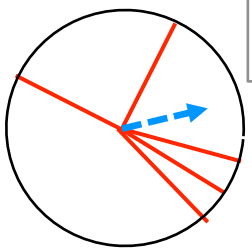


平均は？

$$Re^{i\Theta} = \left\langle e^{i\theta_j} \right\rangle$$

$$Re^{i2\Theta} = \left\langle e^{i2\theta_j} \right\rangle$$

×2



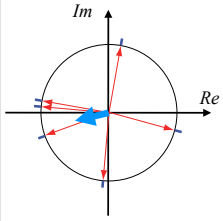
いったん角度を2倍して平均値を求め、それを2で割る

{ 170°,175°, 160°,65°, 35°} 平均値: $\Theta = 6.5°$

↓ ↑

{ 340°,350°, 320°,130°, 70°} 平均値: $2\Theta = 12.9°$

ここまでのまとめ



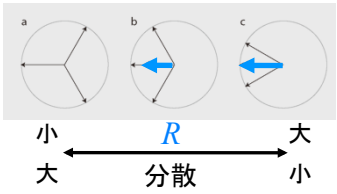
$$(R\cos\Theta, R\sin\Theta) = \frac{1}{N}\left(\sum_j \cos\theta_j, \sum_j \sin\theta_j\right)$$

角度 Θ
長さRのベクトル

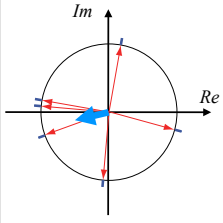
$$Re^{i\Theta} = \frac{1}{N}\sum_j e^{i\theta_j} = \left\langle e^{i\theta_j} \right\rangle$$

複素平面における表記

平均	分散	標準偏差
Θ	$V \equiv 1 - R \quad (0 \le V \le 1)$	$S \equiv \sqrt{-2\log(R)}$



ここまでのまとめ



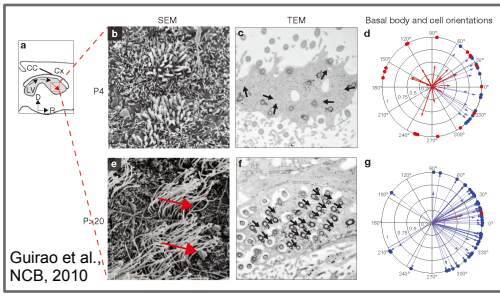
$$(R\cos\Theta, R\sin\Theta) = \frac{1}{N}\left(\sum_j \cos\theta_j, \sum_j \sin\theta_j\right)$$

角度 Θ
長さRのベクトル

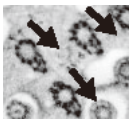
$$Re^{i\Theta} = \frac{1}{N}\sum_j e^{i\theta_j} = \left\langle e^{i\theta_j} \right\rangle$$

複素平面における表記

平均	分散	標準偏差
Θ	$V \equiv 1 - R \quad (0 \le V \le 1)$	$S \equiv \sqrt{-2\log(R)}$

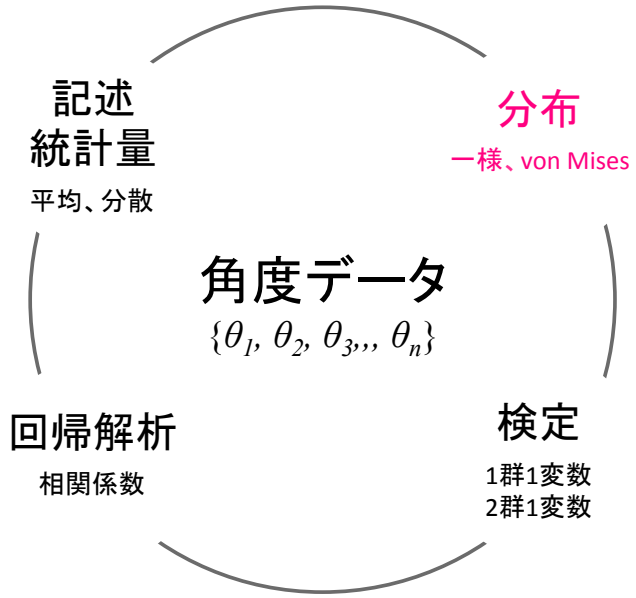


たくさんの繊毛/1細胞
各繊毛のBasal bodyの角度を測定
細胞ごとに角度 Θ 長さRのベクトルを表示
P4（生まれて4日後）とP20で比較



$$R_{P4} < R_{P20}$$
$$V_{P4} > V_{P20}$$

マップ



von Mises分布

方向統計で基本となる分布 （リニア統計での正規分布的な位置づけ）

$$P(\theta) = \frac{\exp\left(\kappa\cos\left(\theta-\mu\right)\right)}{2\pi I_0(\kappa)} \quad \propto \quad \exp\left(\kappa\cos\left(\theta-\mu\right)\right)$$

パラメタ (μ,κ)

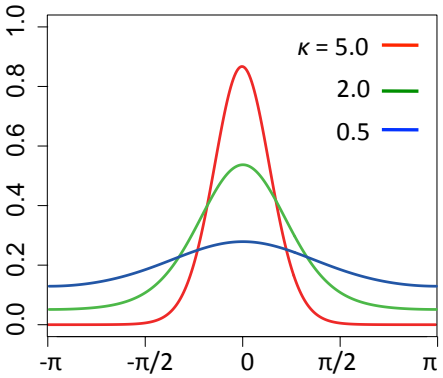
平均 μ

$R = I_1(\kappa)/I_0(\kappa)$

$I_p(\kappa)$はp次変形ベッセル関数

$\kappa \to 0$ 一様分布

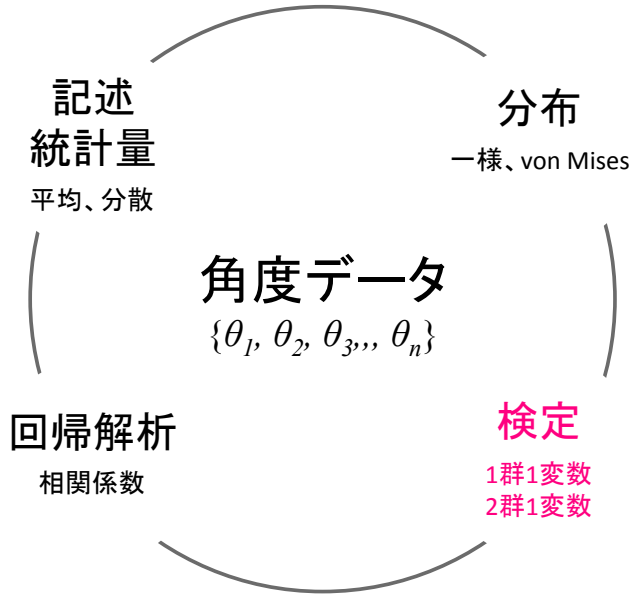
$\kappa \to \infty$ 正規分布



von Mises分布

未発表データのため割愛させていただきます。

マップ



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-Watoson-Wheeler test

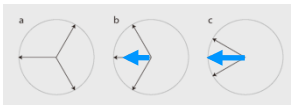
2群のデータは同じ分布に従っているのか？

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　$Re^{i\Theta} = \left\langle e^{i\theta_j} \right\rangle$　　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。

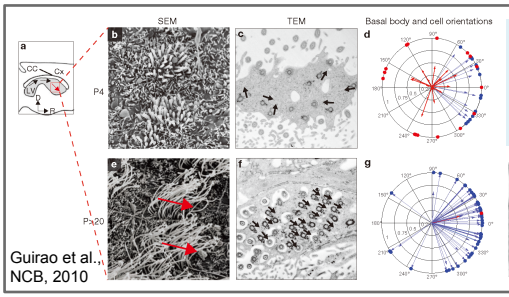


一様分布(帰無仮説)のもとでは、サンプル数nの時に $Z = nR^2$ が出る確率は

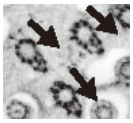
$$P = e^{-Z}\left(1 + \frac{2Z - Z^2}{4n} - \frac{24Z - 132Z^2 + 76Z^3 - 9Z^4}{288n^2}\right) \sim e^{-Z}$$

(p値)　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)



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



$R_{P4} < R_{P20}$

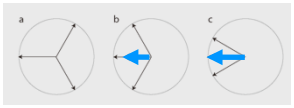
P≥0.05　P<0.05

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　$Re^{i\Theta} = \left\langle e^{i\theta_j} \right\rangle$　　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

$$P = e^{-Z}\left(1 + \frac{2Z - Z^2}{4n} - \frac{24Z - 132Z^2 + 76Z^3 - 9Z^4}{288n^2}\right) \sim e^{-Z}$$

(p値)　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 $R_0 = R\cos(\Theta - \theta_0)$　を計算

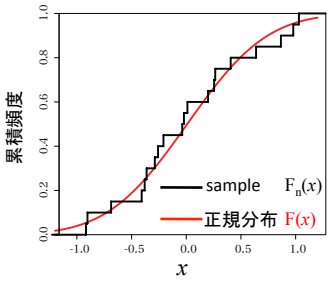
② R_0 が大きければ角度θ_0に偏っている度合いが大きいといえる。

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

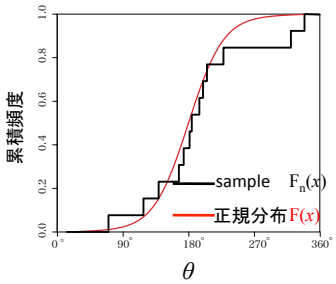
Kolmogorov-Smirnov (KS) 検定の角度データ版
1標本の適合度検定（ある分布に従っているのか？）　従っていないと言えるか？
(2標本が同じ分布からサンプルされているのか？)　異なっていると言えるか？

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



検定量 $D = \max |F_n(x) - F(x)|$

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

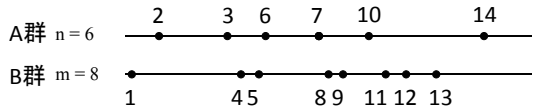


検定量 $V = \max (F_n(x) - F(x))+\max(F(x) - F_n(x))$

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 (リニアデータ)　$\{x_1,x_2,x_3,,,x_n\}$ と $\{y_1,y_2,y_3,,,y_m\}$ → n+m個のデータを混ぜて順位付け
（x: n個, y: m個, $n < m$）



$R = 42$ (A群の順位和)
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-WhitenyのU検定の角度データ版

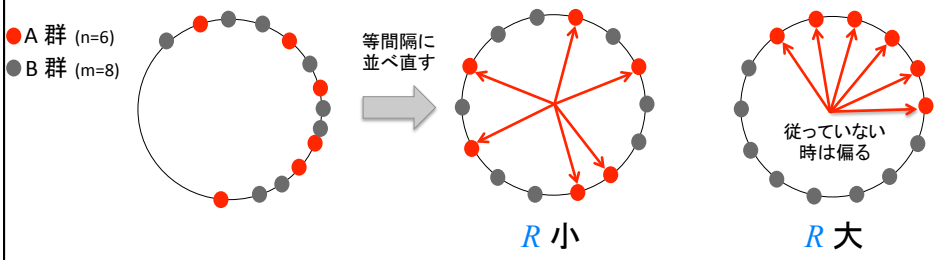
Mardia-Watson-Wheeler test: 2群は同じ分布に従っているか

MWW検定
(角度データ)

$\{\theta_1, \theta_2, \theta_3,,, \theta_n\}$ と $\{\psi_1, \psi_2, \psi_3,,, \psi_m\}$
n個 ＜ m個

→ n+m個のデータを混ぜて小さい順に0-2πで等間隔に並べる

→ $\left\{\frac{2\pi\Theta_1}{n+m}, \frac{2\pi\Theta_2}{n+m}, ,, \frac{2\pi\Theta_n}{n+m}\right\}$ $\left\{\frac{2\pi\Psi_1}{n+m}, \frac{2\pi\Psi_2}{n+m}, ,, \frac{2\pi\Psi_m}{n+m}\right\}$ Θ_j, Ψ_jは 0 から $n+m-1$ の整数

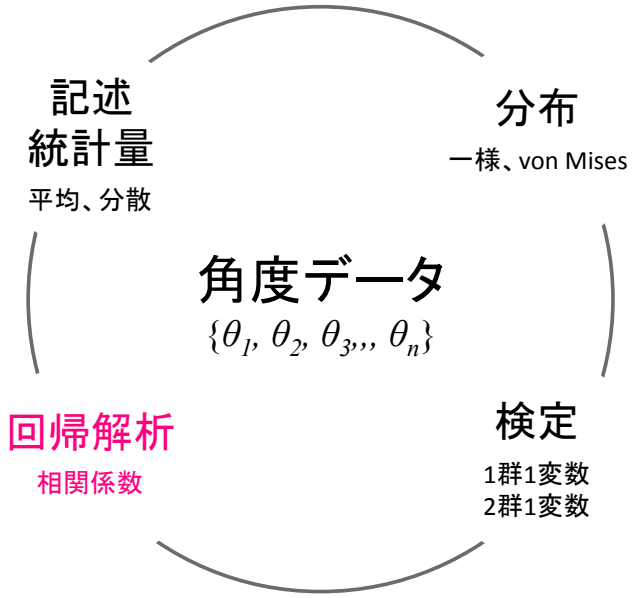


R 小　　R 大

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の大きさで帰無仮説（同じ分布から得られた）を棄却するか否かを判定する

マップ



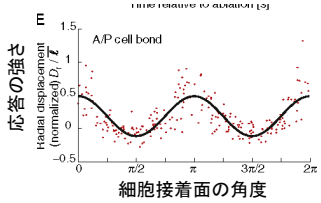
相関・回帰分析

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

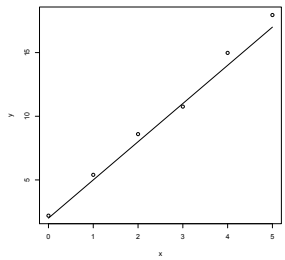
最小二乗法 $L_{ll}(a,b)$, $L_{cl}(a,b,c)$を最小にする(a,b,c)を求める.

リニア-リニアデータ(x_j,y_j)

$$y = a + bx$$

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

bが大きければx依存性が大きい

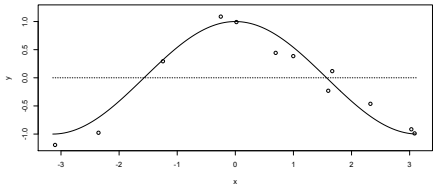


リニア-角度データ (S_j,θ_j)

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

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

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

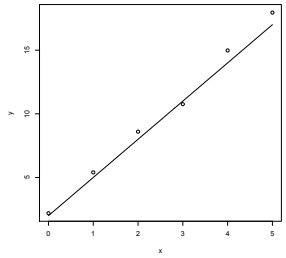
最小二乗法 $L_{ll}(a,b)$, $L_{cl}(a,b,c)$を最小にする(a,b,c)を求める.

リニア-リニアデータ(x_j,y_j)

$$y = a + bx$$

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

bが大きければx依存性が大きい

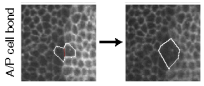


リニア-角度データ (S_j,θ_j)

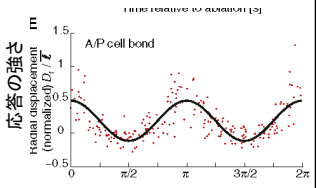
$$S = a + b'\cos(\theta - \mu)$$
$$= a + b\cos\theta + c\sin\theta$$

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

b'が大きければ角度依存性が大きい



Landsberg et al.,
Current Biol, 2009



細胞接着面の角度

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

$$r_{xy} = \left\langle \Delta x_j \Delta y_j \right\rangle \Big/ \sqrt{\left\langle \Delta x_j^2 \right\rangle} \sqrt{\left\langle \Delta y_j^2 \right\rangle}$$

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

$$S = a + b'\cos(\theta - \mu)$$
$$= a + b\cos\theta + c\sin\theta \quad (❖)$$

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

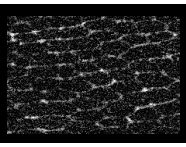
$$\rho_{\theta,S} = \sqrt{\frac{r_{cS}^2 + r_{sS}^2 - 2r_{cS}r_{sS}r_{cs}}{1 - r_{cs}^2}}$$

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

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

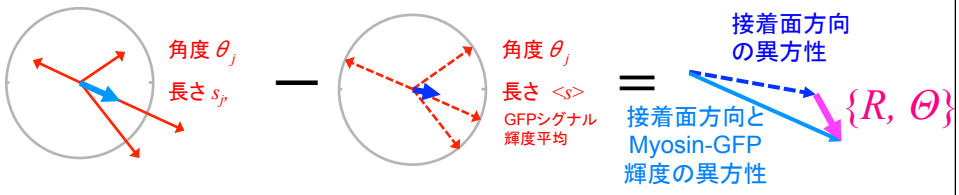
→ *Bootstrap*(詳細は文献参照)

シグナル異方性特徴づけのための統計量

Myosin-GFP 

各接着面について、その角度とMyosin-GFP輝度のペアのデータセット $\{\theta_j, s_j\}$

「接着面方向とMyosin-GFP輝度の異方性」 — 「接着面方向の異方性」 = 「Myosin-GFP輝度の異方性」

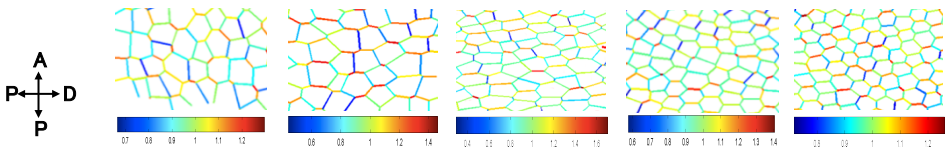


$$\left\langle s_j e^{i2\theta_j} \right\rangle - \left\langle s_j \right\rangle \left\langle e^{i2\theta_j} \right\rangle = R\langle s \rangle e^{i2\Theta}$$

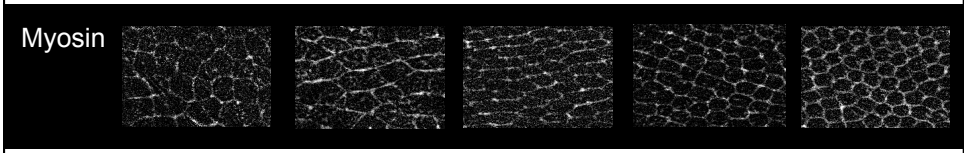
- ベクトルの角度 Θ がMyosin-GFP輝度の異方性の方向
- ベクトルの長さ R が異方性の強さ (s_jとθ_jが独立ならばRは0に近い)

解析結果

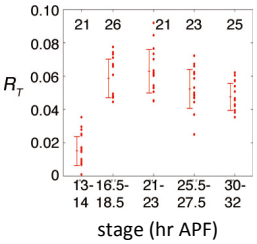
張力



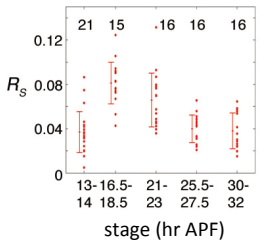
Myosin



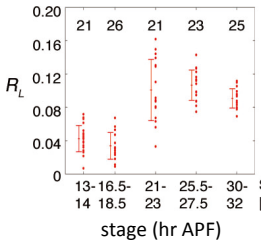
張力の異方性



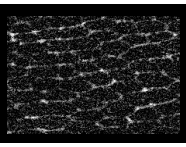
ミオシンの異方性



接着面の長さの異方性



ブートストラップ法による統計的有意性の検証

Myosin-GFP 

各接着面について、
その角度とMyosin-GFP輝度の
ペアのデータセット$\{\theta_j, s_j\}$

「接着面方向とMyosin-GFP輝度の異方性」 — 「接着面方向の異方性」 = 「Myosin-GFP輝度の異方性」 $\{R, \Theta\}$

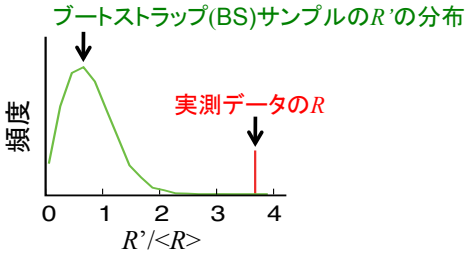
ブートストラップ法

オリジナル(実測)データ $\{\theta_j, s_j\} \rightarrow \{R, \Theta\}$

s_jをランダムに入れ替える

×10000回 (たくさん)

ブートストラップ(BS)サンプル $\{\theta_j, s_j'\} \rightarrow \{R', \Theta'\}$



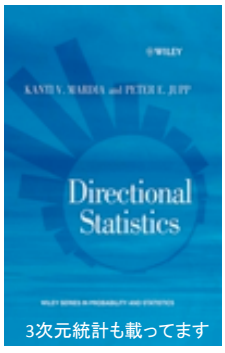
- 実測値のRがBS平均から外れている → 異方性があることを示唆
- *p*-値の評価

Sugimura & Ishihara, 2013

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



3次元統計も載ってます

Batschelet
Circular statistics in biology

絶版

日本語の文献は少ない

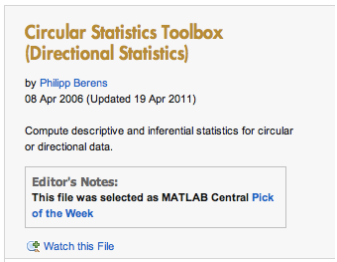
「逆」引き統計学実践統計テスト100(カンジ著、池谷•久我訳)に検定がいくつか載ってます

実装

MATLAB

circular statistics toolbox

by Philippe Berens

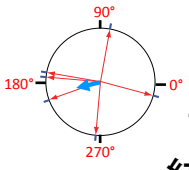


R

circular statistics package

※ 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。

まとめ



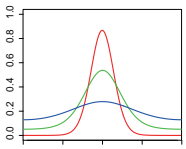
角度 Θ
長さ R

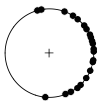
記述統計量

平均、分散

分布

一様、von Mises





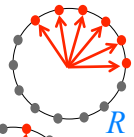
角度データ

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

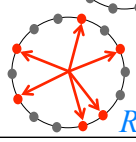
検定

1群1変数

2群1変数



R 大

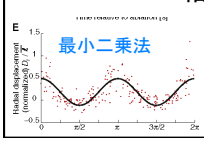


R 小

回帰解析

相関係数

$$\rho_{\theta,s} = \sqrt{\frac{r_{cs}^2 + r_{sc}^2 - 2r_{cs}r_{sc}r_{cc}}{1 - r_{cc}^2}}$$



以下、当日の講演で非表示にしていたスライドです。

代表的な分布

一様分布　$P(\theta)=\frac{1}{2\pi}$　平均 undefined　$R = 0$

コサイン分布　$P(\theta)=\frac{1}{2\pi}\left(1+C\cos(\theta-\mu)\right)$　平均 μ　$R = C/2$

von Mises分布

角度統計で基本となる分布
（リニア統計での正規分布的な位置づけ）

$$P(\theta)=\frac{\exp\left(\kappa\cos(\theta-\mu)\right)}{2\pi I_0(\kappa)}$$

平均 μ　$R = I_1(\kappa)/I_0(\kappa)$

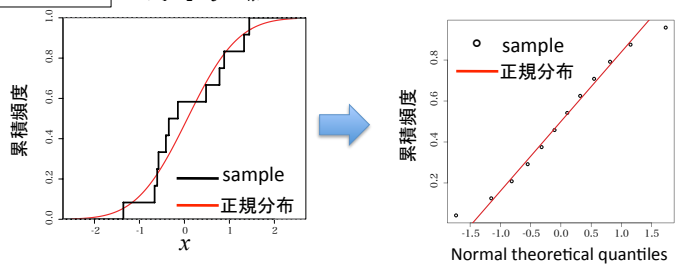
$I_p(\kappa)$はp次変形ベッセル関数

$\kappa \to 0$　一様分布

$\kappa \to \infty$　正規分布

見た目で正規性をチェックしてみる

リニアデータ $\{x_1, x_2, x_3, .., x_n\}$



角度データ $\{\theta_1, \theta_2, \theta_3, .., \theta_n\}$