# 数学を用いる生物学：理念・概念と実践・方法論

## 8月28日

10:00-10:30 島谷健一郎(統数研)開催挨拶, presence-only dataとポアソン点過程
10:30-11:50 大泉嶺(人口問題研究所)多地域レスリー行列の理論と応用～日本の人口減少における国内・国際移動の影響～
12:40-14:00 佐竹曉子(九州大) 大規模同調開花は将来どうなるか？長期データを用いた将来予測
14:20-15:40岩田繁英(東京海洋大)空間構造を有する生物資源に対する漁獲枠の決定方法に関する考察
16:00-16:40 青木聡志(環境研)遺伝的相関の総和を最小にする生物個体の空間抽出とその下での遺伝的多様性
16:40-17:20 井上巨人(神戸大)・石原孝・武沢幸雄・東口信行・宮地麻実・福山千夏(AQUARIUM x ART á toa)・久野宗郎(ASICS)歩数データから読み解くカメの個性
17:20- 佐藤雄亮(東京大農)分布拡大条件の繁殖様式による違い：繁殖保証と移住荷重
本間千夏(秋田県立大)岩手県の落葉樹林における種子生産量の長期パターン：個体群レベル・個体ベースの推定

## 8月29日

10:00-10:30 島谷健一郎(統数研)散布制限を伴う植物個体群動態の空間点過程モデル
10:30-11:50 後藤佑介(名古屋大)アホウドリの風に対する移動戦略
12:50-14:10香川幸太郎(遺伝研)プログラミングで進化の法則を探る：個体ベース・モデルによる進化の理論研究
14:30-15:10 深澤圭太(環境研)空間標識再捕獲法による個体密度と景観連結性の推定
15:10-15:50 高須夫悟(奈良女子大)空間個体群動態の数理：個体ベースの出生・死亡・空間移動を数理的に記述する方法について
15:50-16:30 阪上雅昭(京都大)変分オートエンコーダによる乳幼児の表出語彙数発達の解析
16:30-17:00 討論その他

presence-only data

・そこに この種がいた・みつけた

・動物のGPS軌跡

presence-absence data

・Sampling plots における在・不在 (0-1) データ

・Sampling plots における個体数 (abundance) データ

・監視カメラなど：写っていた (presence) 情報だけでなく写っていなかった (absence) 情報を活用できるかが鍵

環境情報 $l_i$ も取得し（野外調査、GIS, …）、得られたデータを総動員して調査していない（環境情報はある）区域も含めて species distribution model

$d(\mathbf{x}; \boldsymbol{\theta}) = \exp(\theta_0 + \theta_1 l_1(\mathbf{x}) + \cdots + \theta_m l_m(\mathbf{x}))$ + ランダム効果を作る

$d(\mathbf{x};\boldsymbol{\theta}) = \exp(\theta_0 + \theta_1 l_1(\mathbf{x}) + \cdots + \theta_m l_m(\mathbf{x}))$ は何を表すか？

2020年ころ、Gelfand などがマジメに論文で議論
その種の個体がいる確率？

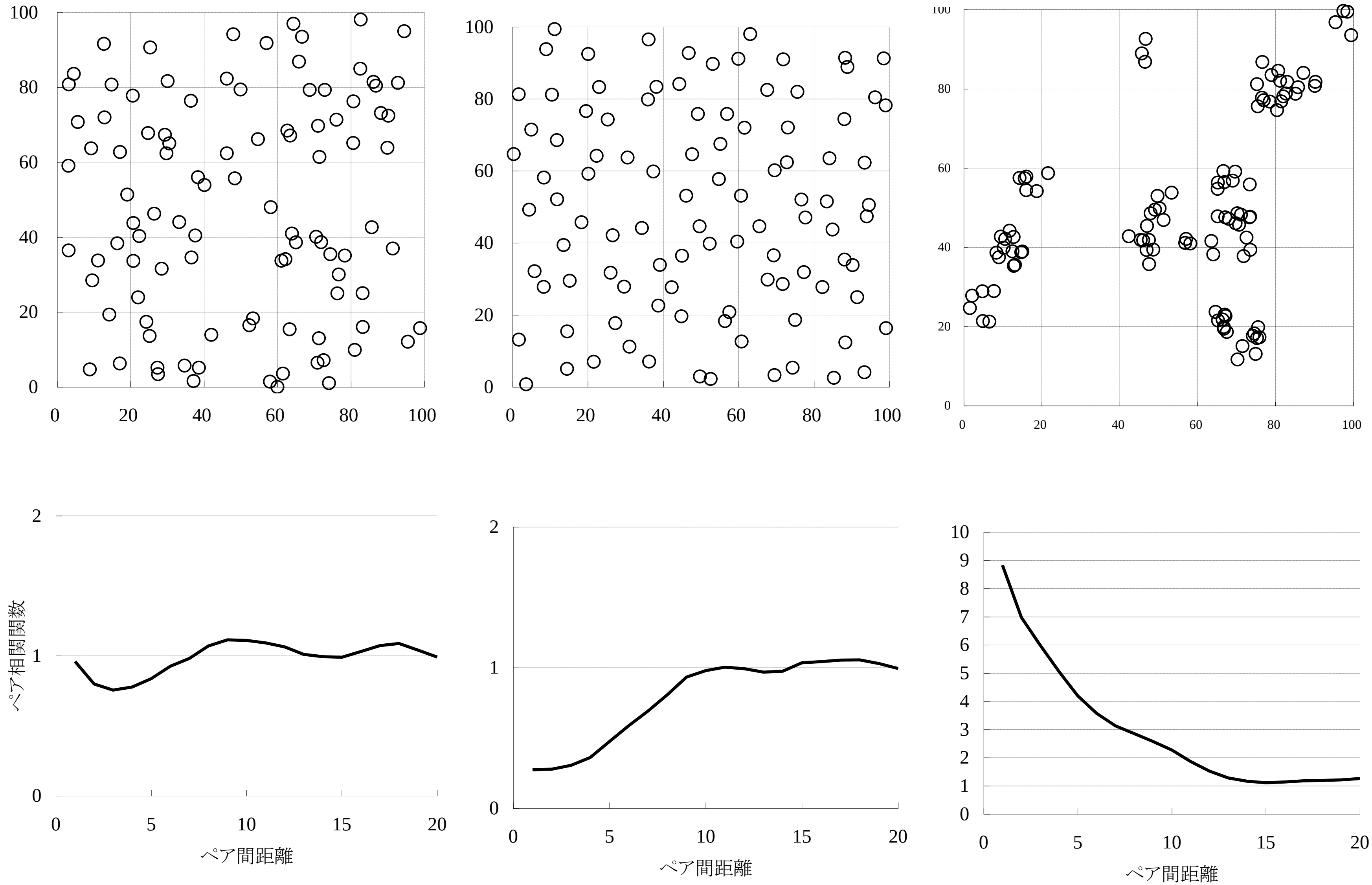
**空間点過程モデルの強度関数という数学の枠組みが一番適している**
**x: のまわりの微小単位面積にその種の個体がいる確率**
領域 A で $d(\mathbf{x};\theta)$ を積分すると個体数（の期待値）になる

現在の問題

1. 実際のデータと $d(\mathbf{x};\theta)$ をつなげるには、例えば、調査努力関数 $effort(\mathbf{x};\theta)$ と発見率関数 $detection(\mathbf{x};\theta)$ も必要.（状態空間モデルにおける）観察モデルの構築と推定法.
2. 実際の種分布は非定常ポアソン過程では表せない（相互作用による集中あるいは距離を隔てる分布）.より実態に近い空間点過程モデルと未知パラメータのデータからの推定法の開発.

# 点パターンの特徴を表す統計量：ペア相関関数：
# ある距離離れたペア数（を正規化）

ランダム　　レギュラー　　集中



# 点パターンの特徴を表す統計量：ペア相関関数：
# ある距離離れたペア数（を正規化）

ランダム　　　　レギュラー　　　　集中

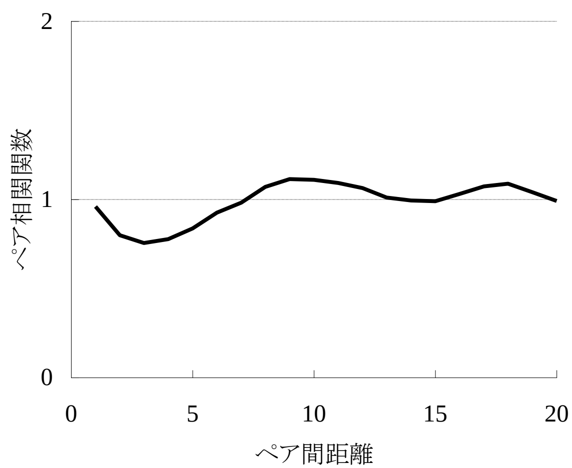
1st order moment = intensity function **強度関数**

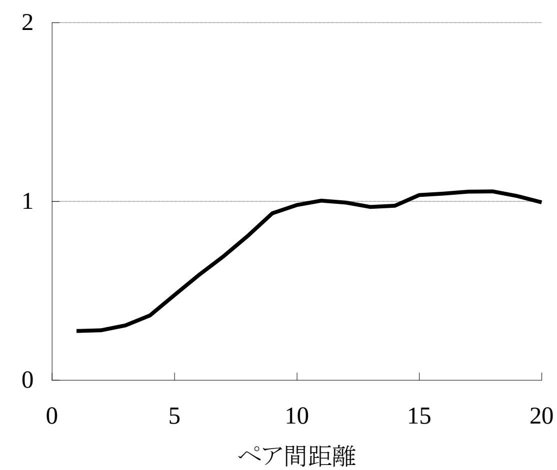
$\rho^{(1)}(x)$:  $x$ **を重心とする無限小単位面積に個体がいる確率（定常なら平均密度）**

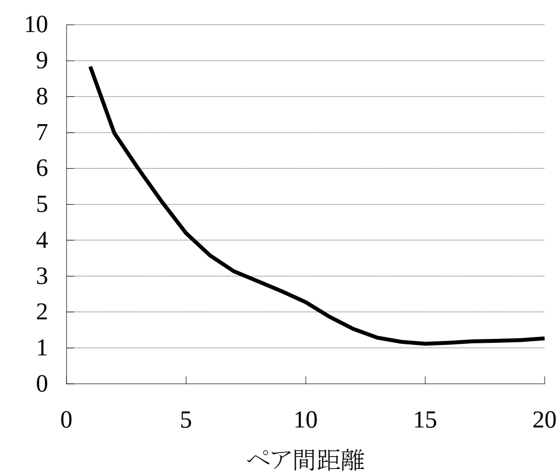
2nd order moment

$\rho^{(2)}(x, y)$: $x, y$**を重心とする2つの無限小単位面積の両方に個体がいる確率**

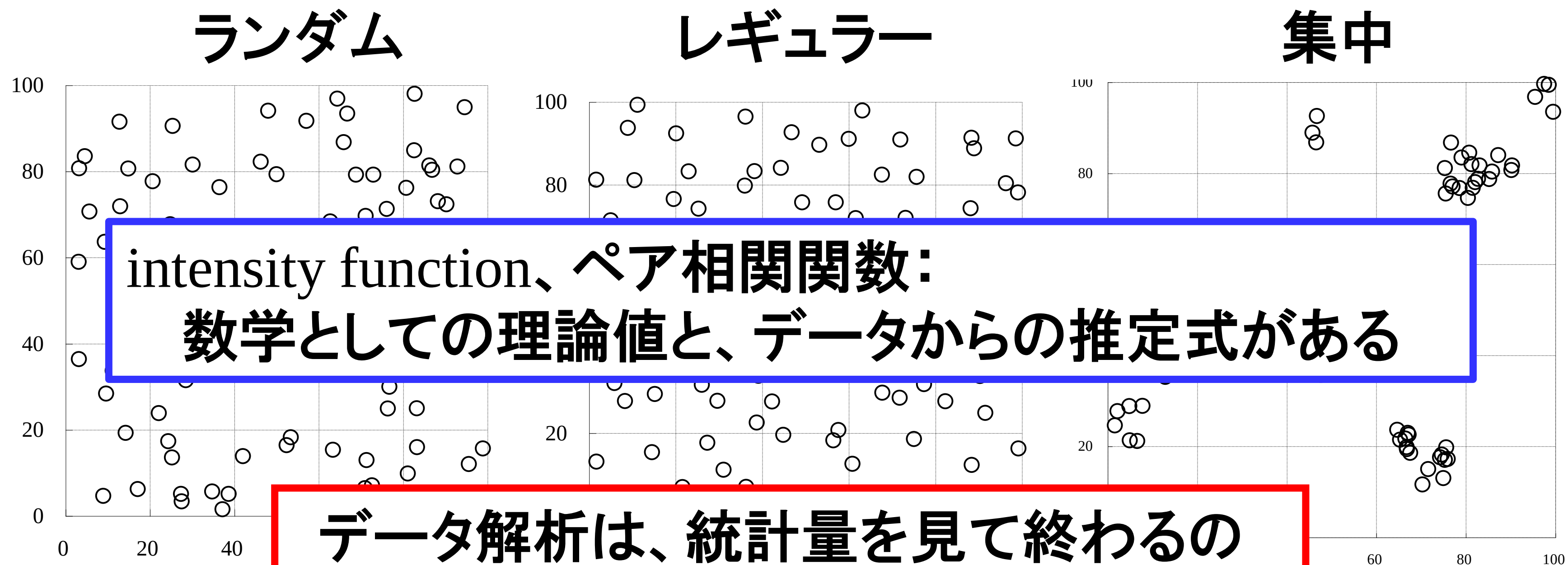
**ペア相関関数：** $\rho^{(2)}(x, y)$ **を** $\rho^{(1)}(x)\rho^{(1)}(y)$ **で割って正規化**



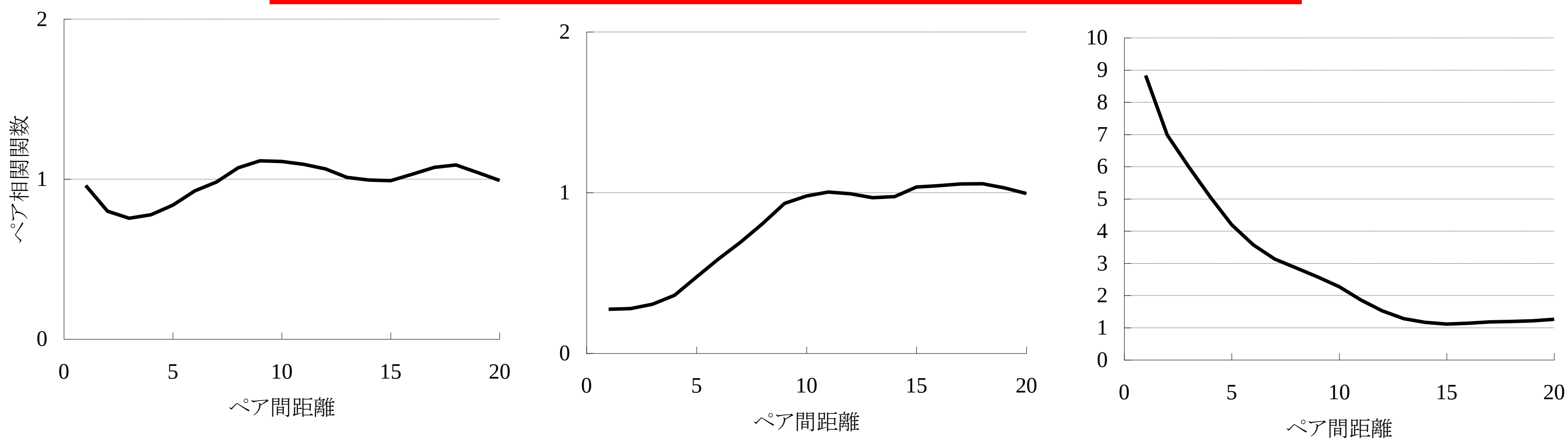



# 点パターンの特徴を表す統計量：ペア相関関数：
# ある距離離れたペア数（を正規化）

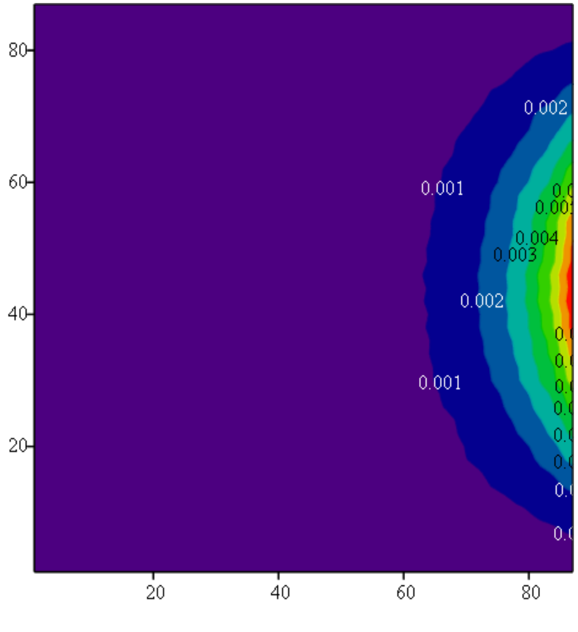



**intensity function、ペア相関関数：**
**数学としての理論値と、データからの推定式がある**

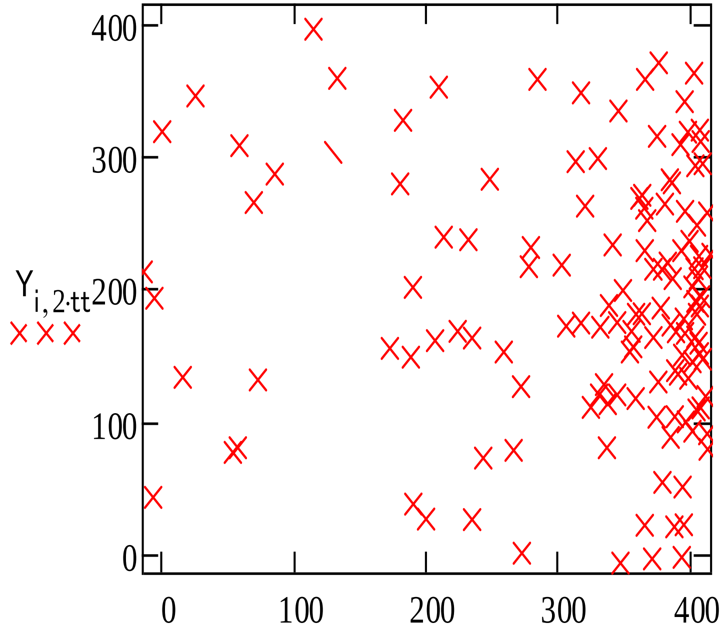
**データ解析は、統計量を見て終わるのでなく、そこから始まる（＝モデリング）**



# 非定常ポアソン過程




環境要因の変異



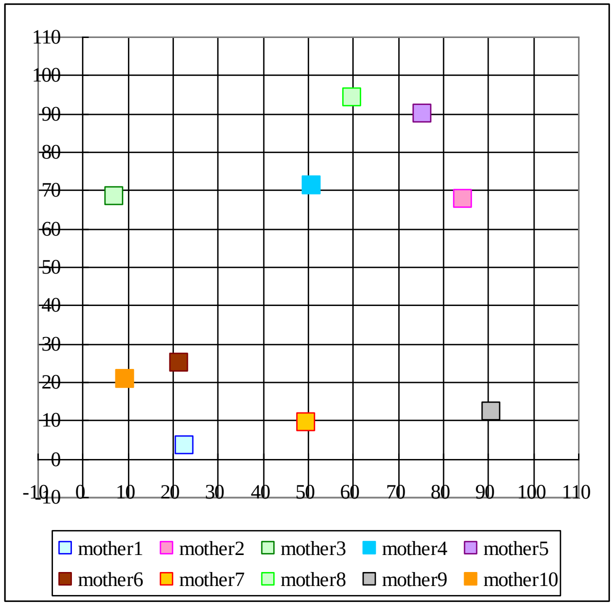

点配置例

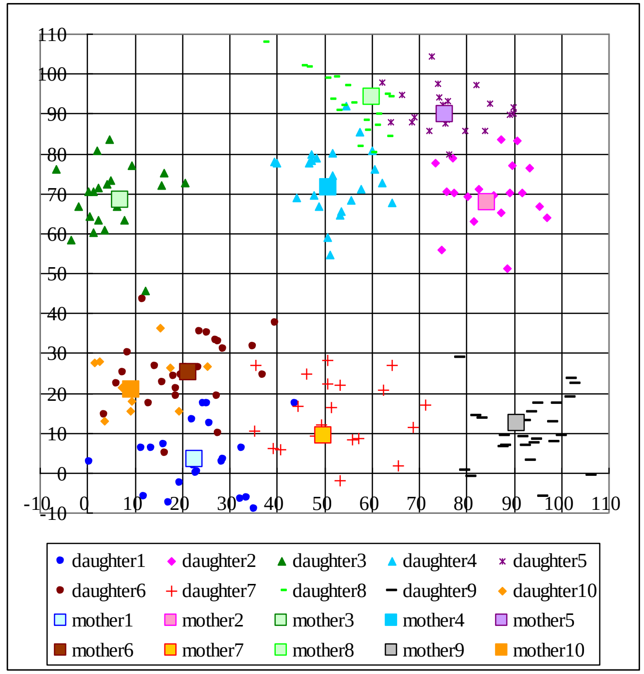
・強度関数 $d(\mathbf{x};\,\boldsymbol{\theta}) = \exp(\theta_0 + \theta_1 l_1(\mathbf{x}) + \cdots + \theta_m l_m(\mathbf{x}))$

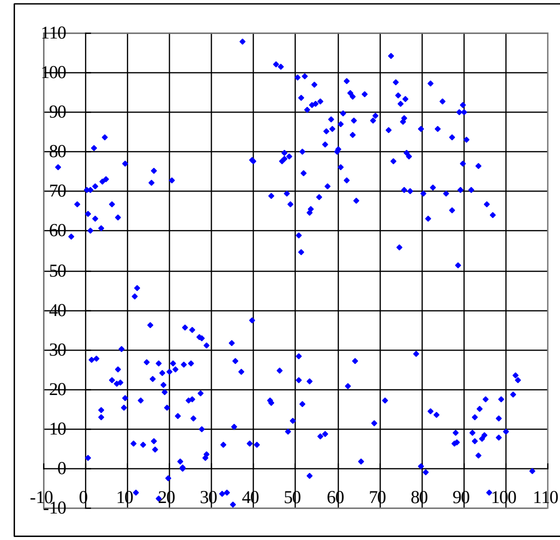
$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$:調査プロット $A$ の中の点配置データ

尤度関数 $$L(\theta) = \underbrace{\prod_{i=1}^{n} d(\mathbf{x}_i;\theta)}_{\mathbf{x}_1,\,\mathbf{x}_2,\,\ldots,\,\mathbf{x}_n\text{にあるという情報}} \cdot \underbrace{e^{-\int_A d(\mathbf{x};\theta)d\mathbf{x}}}_{\mathbf{x}_1,\,\mathbf{x}_2,\,\ldots,\,\mathbf{x}_n\text{以外にはないという情報}}$$

# 基本モデル2: 定常ネイマン･スコット (Thomas) 過程



1. 母個体はランダムに密度 $d$ で分布

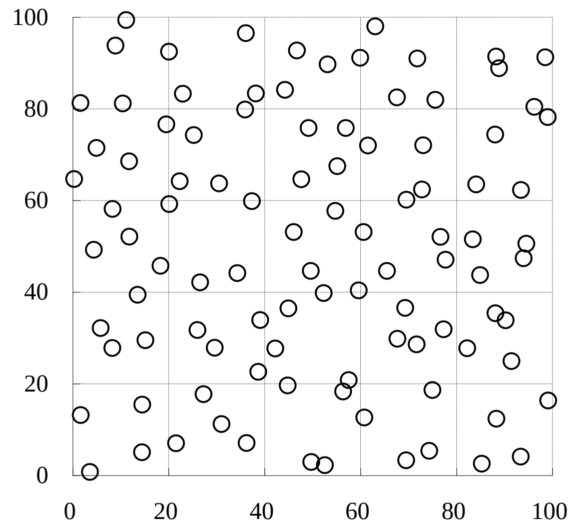
2. 各母個体は強度 $U$ のポアソン分布に従う$u$ 個の娘を生む。

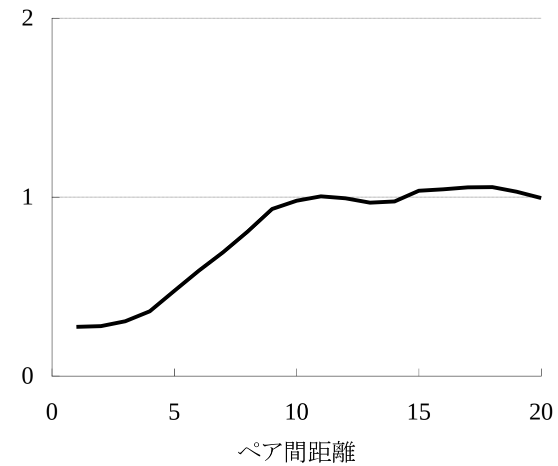
3. 娘は2次元正規分布の密度関数 $d_{Nl}(r;\sigma^2)$ に従って散布

4. 母は死亡し娘だけが残る

# 基本モデル3: Gibbs点過程（競合で近接ペアは少ない）

## レギュラー

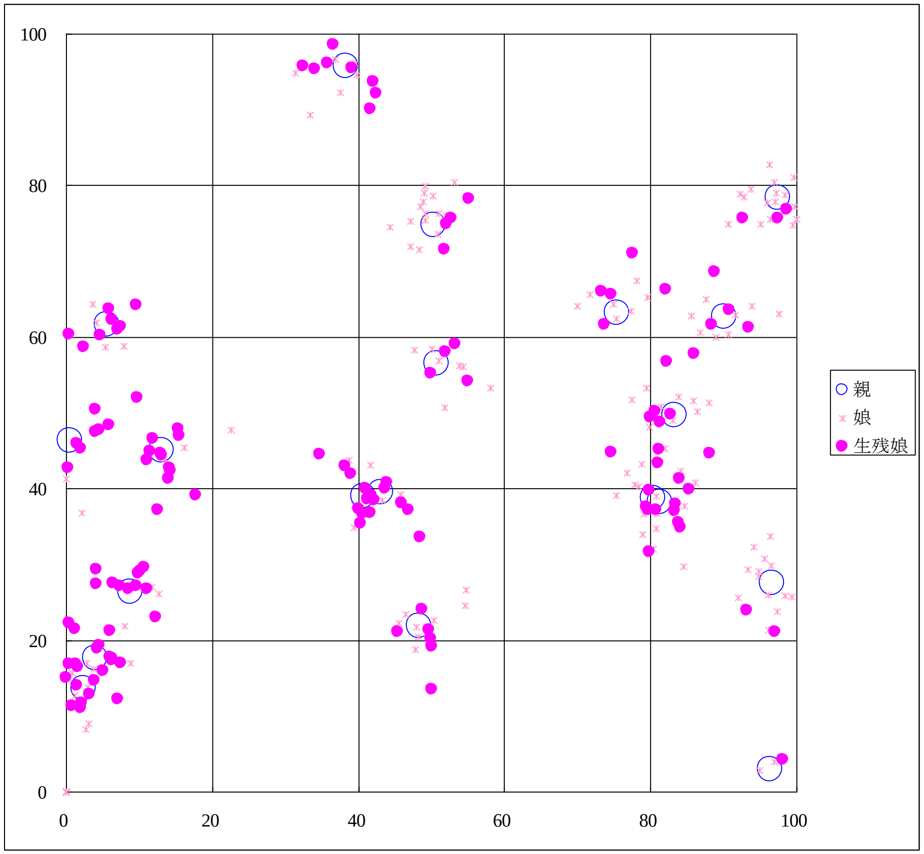


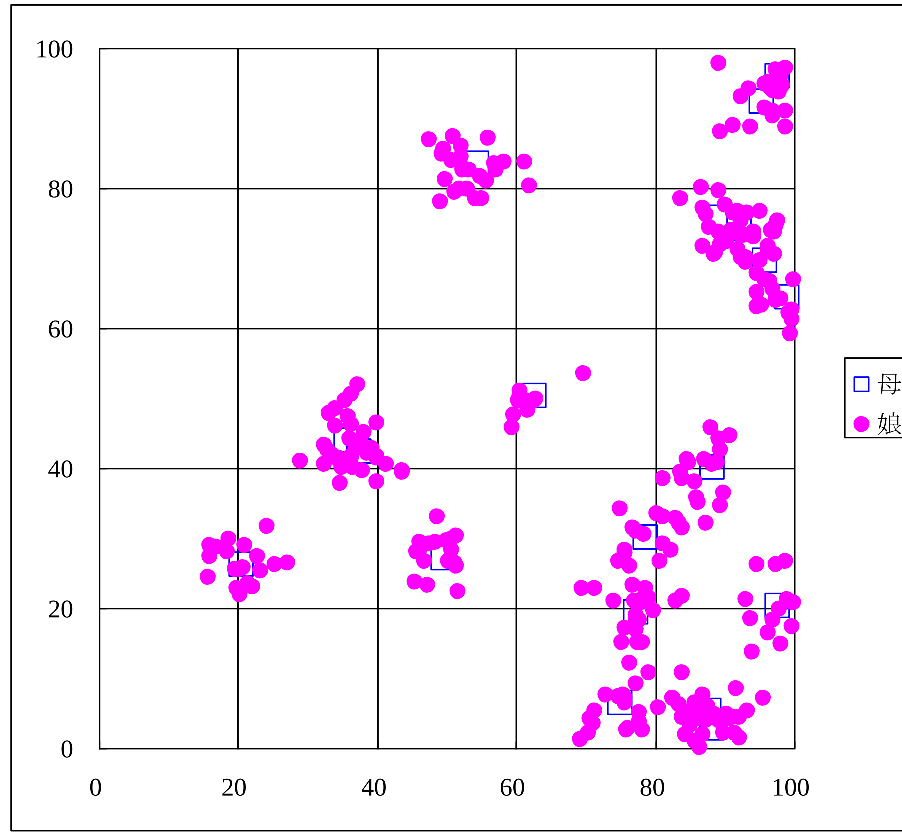
2. **植物個体群に必要な空間点過程：繰り返し（繁殖）と非定常性（環境依存）**

# 非定常ポアソン過程とネイマン・スコット過程を結合する

## 娘個体に環境依存生残率




## 母親密度が環境依存

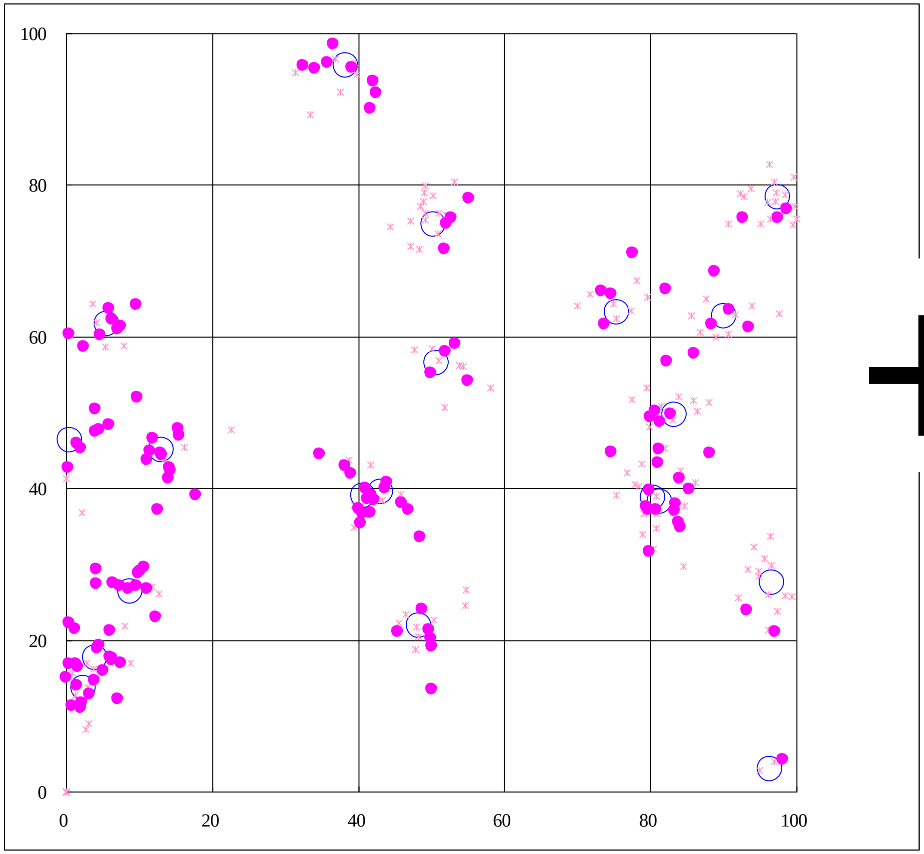



## パラメータ推定法：いろいろ開発済み

2. 植物個体群に必要な空間点過程：繰り返し（繁殖）と非定常性（環境依存）

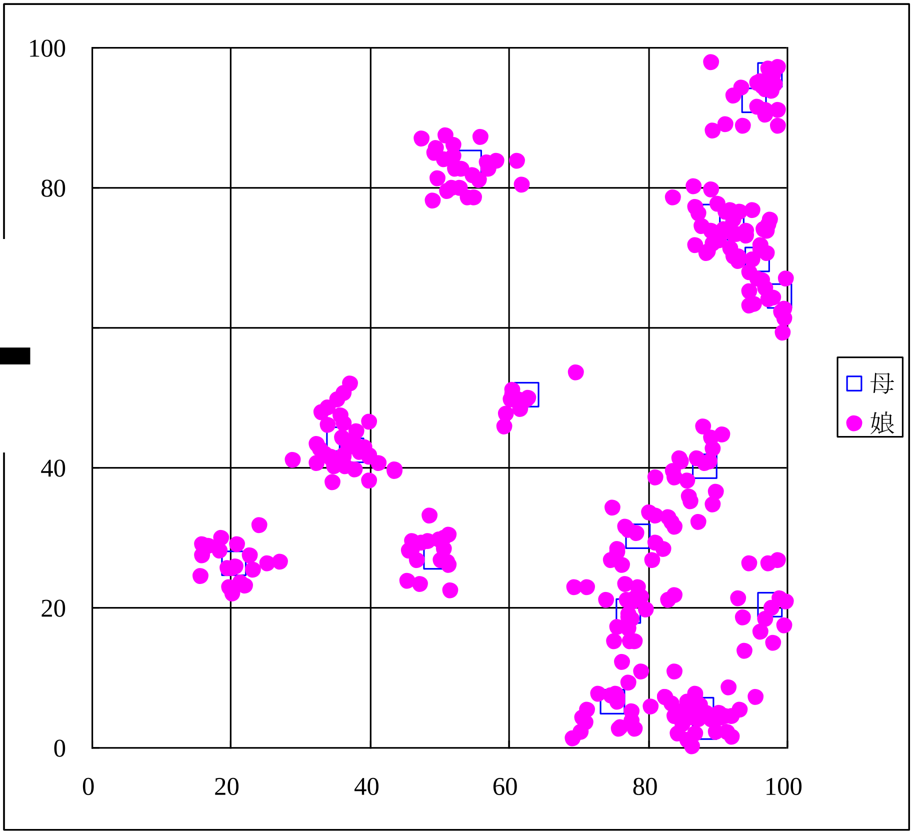
# 非定常ポアソン過程とネイマン・スコット過程を結合する

## 娘個体に環境依存生残率




＋

## 母親密度が環境依存




強度関数やペア相関関数の理論値・データからの推定式

最尤法、ベイズ推定、モーメント法、…でパラメータ推定

・母親個体群： 1次モーメント： $\rho_M^{(1)}(x)$ 2次モーメント： $\rho_M^{(2)}(x, y)$
・娘個体群： 1次モーメント： $\rho_D^{(1)}(x)$ 2次モーメント： $\rho_D^{(2)}(x, y)$

# 漸化式

$$\rho_D^{(1)}(x) = \int_A \rho_M^{(1)}(z)\mathbf{E}(U)d_{Nl}(x; z, r^2)dz \cdot s(x)$$

$A$：親が存在しうる領域（無限平面）

$z$ に親がいて $U$ 個の娘を産み $x$ に散布され確率 $s(x)$で生残

$$\rho_D^{(2)}(x, y) = \rho_D^{(1)}(x)\rho_D^{(1)}(y)$$

$x$ にいる娘と $y$ にいる娘の母親は別

$$+\int_A\int_A \rho_M^{(1)}(z)\mathbf{E}(U(U-1))d_{Nl}(x; z, r^2)d_{Nl}(y; z, r^2)dz \cdot s(x)s(y)$$

$z$ に共通の母親がいて $U$ 個の娘を産み $x$, $y$ に散布され確率 $s(x)$, $s(y)$ で生残

$$\rho_M^{(1)}(x; \boldsymbol{\theta}) = \exp(\theta_0 + \theta_1 l_1(x) + \cdots + \theta_m l_m(x))$$

1次モーメントは線形回帰の形にし(+link関数)、環境因子 $i$ は、サンプル地点 $v_{1,\mathrm{i}}$, …, $v_{K,i}$ での数値 $V_{1,\mathrm{i}}$, …, $V_{K,i}$ を正規分布カーネルなどで平滑化して使うことが多い。

$$l_i(x) = \sum_k d_{Nl}(x; v_{k,i}, \sigma^2)V_{k,i}$$

だから積分を初等関数で表せない。

・母親個体群： １次モーメント： $\rho_M^{(1)}(x)$　２次モーメント： $\rho_M^{(2)}(x, y)$
・娘個体群：　１次モーメント： $\rho_D^{(1)}(x)$　２次モーメント： $\rho_D^{(2)}(x, y)$

# 漸化式

$$\rho_D^{(1)}(x) = \int_A \rho_M^{(1)}(z)\mathbf{E}(U)d_{Nl}(x; z, r^2)dz \cdot s(x)$$

$A$：親が存在しうる領域（無限平面）

$z$ に親がいて $U$ 個の娘を産み $x$ に散布され確率 $s(x)$で生残

$$\rho_D^{(2)}(x, y) = \rho_D^{(1)}(x)\rho_D^{(1)}(y)$$

$x$ にいる娘と $y$ にいる娘の母親は別

$$+\int_A\int_A \rho_M^{(1)}(z)\mathbf{E}(U(U-1))d_{Nl}(x; z, r^2)d_{Nl}(y; z, r^2)dz \cdot s(x)s(y)$$

$z$ に共通の母親がいて $U$ 個の娘を産み $x$, $y$ に散布され確率 $s(x)$, $s(y)$ で生残

$$\rho_M^{(1)}(x; \boldsymbol{\theta}) = \exp(\theta_0 + \theta_1 l_1(x) + \cdots + \theta_m l_m(x))$$

1次モーメントは線形回帰の形にし(+link関数)、環境因子 $i$ は、サンプル地点 $v_{1,i}$, …, $v_{K,i}$ での数値 $V_{1,i}$, …, $V_{K,i}$ を正規分布カーネルなどで平滑化して使うことが多い。

$$l_i(x) = \sum_k d_{Nl}(x; v_{k,i}, \sigma^2)V_{k,i}$$

だから積分を初等関数で表せない。

## 発想を変えて、強度関数をカーネル平滑化する

$$\bar{\rho}_M^{(1)}(x) = \exp(\theta_0 + \sum \theta_i \sum_k d_{Nl}(x; v_{k,i}, \sigma^2) V_{k,i})$$

$$\approx \sum_k d_{Nl}(x; v_{k,i}, \sigma^2) \exp(\theta_0 + \sum \theta_i V_{k,i})$$

# 漸化式

$$\rho_D^{(1)}(x) = \int_A \rho_M^{(1)}(z) \mathbf{E}(U) d_{Nl}(x; z, r^2) dz \cdot s(x)$$

$z$ に親がいて $U$ 個の娘を産み $x$ に散布され確率 $s(x)$ で生残

$$\rho_D^{(2)}(x, y) = \rho_D^{(1)}(x) \rho_D^{(1)}(y)$$
$$+ \int_A \int_A \rho_M^{(1)}(z) \mathbf{E}(U(U-1)) d_{Nl}(x; z, r^2) d_{Nl}(y; z, r^2) dz \cdot s(x) s(y)$$

$$\boxed{\int d_{Nl}(x; w, s^2) d_{Nl}(x; z, t^2) dz = d_{Nl}(w; z, s^2 + t^2)}$$

Convolution formulaにより積分はすべて正規分布で表される

$$\rho_M^{(1)}(x; \boldsymbol{\theta}) = \exp(\theta_0 + \theta_1 l_1(x) + \cdots + \theta_m l_m(x))$$

1次モーメントは線形回帰の形にし(+link関数)、環境因子 $i$ は、サンプル地点 $v_{1,i}$, …, $v_{K,i}$ での数値 $V_{1,i}$, …, $V_{K,i}$ を正規分布カーネルなどで平滑化して使うことが多い。

$$l_i(x) = \sum_k d_{Nl}(x; v_{k,i}, \sigma^2) V_{k,i}$$

だから積分を初等関数で表せない。

・繁殖が繰り返されると項は1つずつ増えていくが基本的に同じ形なので繰り返し可能
・2次モーメントが初等関数で書けるので、composite likelihood, Palm likelihoodなどでデータから未知パラメータを推定可能

# 漸化式

$$\rho_D^{(1)}(x) = \int_A \rho_M^{(1)}(z)\mathbf{E}(U)d_{Nl}(x;z,r^2)dz \cdot s(x)$$

$z$ に親がいて $U$ 個の娘を産み $x$ に散布され確率 $s(x)$で生残

$$\rho_D^{(2)}(x,y) = \rho_D^{(1)}(x)\rho_D^{(1)}(y) + \int_A\int_A \rho_M^{(1)}(z)\mathbf{E}(U(U-1))d_{Nl}(x;z,r^2)d_{Nl}(y;z,r^2)dz \cdot s(x)s(y)$$

$$\boxed{\int d_{Nl}(x;w,s^2)d_{Nl}(x;z,t^2)dz = d_{Nl}(w;z,s^2+t^2)}$$

Convolution formulaにより積分はすべて正規分布で表される

## 祖母が定常ポアソンのときの2代目の2次モーメントの形

$$\rho_D^{(2)}(x,y)$$
$$= \rho_D^{(1)}(x)\rho_D^{(1)}(y)$$ 母の違う他人ペア
$$+ d_{Nl}(x;y,S_1^2)\sum_k C_1^x d_{Nl}(x,v_k,T_k^2)C_1^y d_{Nl}(y,v_k,T_k^2)$$ 母が同じ姉妹ペア
$$+ d_{Nl}(x;y,S_2^2)\sum_h C_2^x d_{Nl}(x,u_h,R_h^2)C_2^y d_{Nl}(y,u_h,R_h^2)$$ 祖母が同じ従姉妹ペア

・モデルのアイデアは素朴かつ生態学的に理に叶っている(単純過ぎる?)
・シミュレーションで実現させるのも容易

（空間）点過程で難しいのは

1. 実データから未知パラメータを推定する統計手法
2. 平衡解の存在や導出のための数学

一般化線形モデル（GLM）やGLMを組み合わせた階層ベイズモデルとは異質の難所…

# 統計モデル：データは（未知の）確率分布からのランダムなサンプルと仮定する

データ　$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$　$n$: サンプルサイズ

**線形回帰モデル**

$$y \mid x \sim Normal(ax+b, \sigma^2)$$

$$y = ax + b + e$$

$$e \sim N(0, \sigma^2)$$

**連続型確率分布**

**尤度関数**

$$L(a,b,\sigma^2) = \frac{e^{-\frac{(y_1-(ax_1+b))^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \times \cdots \times \frac{e^{-\frac{(y_n-(ax_n+b))^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} = \prod_{i=1}^{n} \frac{e^{-\frac{(y_i-(ax_i+b))^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

**ロジスティック回帰**

$$y \mid x \sim Bernoulli(1/\{1+\exp(ax+b)\})$$

**離散型確率分布**

**尤度関数**

$$L(a,b) = \prod_{i=1}^{n} \left(\frac{1}{1+e^{ax_1+b}}\right)^{y_1} \left(1-\frac{1}{1+e^{ax_1+b}}\right)^{1-y_1}$$

# 空間点過程　データは点配置

1. $n$ 個の点の位置（座標）というデータは独立でない
（(サンプル数は1、$n$ はデータの一部）

点過程：$n$ 個の点の位置だから離散的?
点の位置（座標）は連続的?

2. 確率分布：離散と連続が混じっている

点配置の集合＝{$\phi$}U{1個の点の配置}U{2個の点の配置}U…

確率を与える対象(点配置）が実数(ユークリッド空間）や$\{0, 1\}$と根本的に違う

通常の統計学より抽象的な数学が必要

生物系に点過程モデルは無理なのか？

豊富な失敗経験を経ないと
ホンモノの専門にならない

専門性
大学時代(20代？)に(つまらない)繰り返し練習を積む
・正しそうに聞こえる教育論
　若い10代(20代？)のころは好きなことを思い切りやる。
　熱意を持てないことは、10代20代のときに詰め込むのでなく、30代以降に学習すればいい。
・30代以降でも新しい技術の習得は可能だが、それは
　・別な専門性がある
　・共同研究者に教われる　　場合など(島谷仮説?)

そのとき、
・10－20代にイヤイヤ基礎訓練を積んだ
・本気になって関わったことも少しあった
・「かする」くらいだけど、ないわけではなかった
こんな経験が案外と役に立つ

# 統計学（島谷版の全体像）

## 2. データを観る工夫

グラフ、表、集約、特徴量

データを自在に加工して図示

## 1. 統計モデルによる予測・推定

基礎

尤度、最適化、乱数生成、
シミュレーション、…
一般化線形モデル

ギャップ

先端

情報量規準
ベイズ統計
機械学習
…

統計解析の結果(output)を観る

統計学や数理生物学が専門でない人はどの程度に学習しておくか?
数理生物や統計が専門の人はどの程度に生物学を学習しておくか?

# 統計学（島谷版の全体像）

## 2. データを観る工夫

グラフ、表、集約、
特徴量

データを自在に加
工して図示

## 1. 統計モデルによる予測・推定

基礎

ギャップ

先端

尤度、最適化、乱数生成、
シミュレーション、…
一般化線形モデル

情報量規準
ベイズ統計
機械学習
…

統計解析の結果
(output)を観る

当面、一般化線形モデルを(つまらない)統計学の授業で学習した次は、
曼荼羅を見ながら少しの無理をしてギャップを飛び越える…

データ $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ $n$: サンプルサイズ

**線形回帰モデル** $y \mid x \sim Normal(ax+b, \sigma^2)$

$$y = ax + b + e$$

$$e \sim N(0, \sigma^2)$$

**尤度関数**

$$L(a,b,\sigma^2) = \frac{e^{-\frac{(y_1-(ax_1+b))^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \times \cdots \times \frac{e^{-\frac{(y_n-(ax_n+b))^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} = \prod_{i=1}^{n} \frac{e^{-\frac{(y_i-(ax_i+b))^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

**ロジスティック回帰** $y \mid x \sim Bernoulli(1/\{1+\exp(ax+b)\})$

**尤度関数**

$$L(a,b) = \prod_{i=1}^{n} \left(\frac{1}{1+e^{ax_1+b}}\right)^{y_1} \left(1 - \frac{1}{1+e^{ax_1+b}}\right)^{1-y_1}$$

# （空間）点過程

**一般化線形モデルとの(決して多くない)違いを認識できれば、少なくとも共同研究は可能**

# 基礎と発展のギャップの例

## 1. 統計モデルによる予測・推定

ポアソン分布

$$P(X = k) = e^{-d} d^{k} / k!$$

非定常ポアソン点過程モデルの尤度関数

$$L(\theta) = \prod_{i=1}^{n} d(\mathbf{x}_i;\theta) \cdot e^{-\int_A d(\mathbf{x};\theta)\, d\mathbf{x}}$$

- Presence-only data
- 空間点過程を基盤に置く

Species distribution model

基礎

ギャップ

先端

尤度、最適化、乱数生成、
シミュレーション、…
一般化線形モデル

情報量規準
ベイズ統計
機械学習
…

当面、一般化線形モデルを(つまらない)統計学の授業で学習した次は、
曼荼羅を見ながら少しの無理をしてギャップを飛び越える…

# 数学を用いる生物学：理念・概念と実践・方法論

8月28日

今日もまた、様々な曼荼羅を見ていってください。
自分と相性のいい曼荼羅をみつけましょう

10:00-10:30 島

10:30-11:50 大 … 内・国際移動の影響～

12:40-14:00 佐竹暁子（九州大） 大規模同調開花は将来どうなるか？長期データを用いた将来予測

14:20-15:40岩田繁英（東京海洋大）空間構造を有する生物資源に対する漁獲枠の決定方法に関する考察

16:00-16:40 青木聡志（環境研）遺伝的相関の総和を最小にする生物個体の空間抽出とその下での遺伝的多様性

16:40-17:20 井上巨人（神戸大）・石原孝・武沢幸雄・東口信行・宮地麻実・福山千夏（AQUARIUM x ART á toa）・久野宗郎（ASICS）歩数データから読み解くカメの個性

17:20- 佐藤雄亮（東京大農）分布拡大条件の繁殖様式による違い：繁殖保証と移住荷重

本間千夏（秋田県立大）岩手県の落葉樹林における種子生産量の長期パターン：個体群レベル・個体ベースの推定

8月29日

10:00-10:30 島谷健一郎（統数研）散布制限を伴う植物個体群動態の空間点過程モデル

10:30-11:50 後藤佑介（名古屋大）アホウドリの風に対する移動戦略

12:50-14:10香川幸太郎（遺伝研）プログラミングで進化の法則を探る：個体ベース・モデルによる進化の理論研究

14:30-15:10 深澤圭太（環境研）空間標識再捕獲法による個体密度と景観連結性の推定

15:10-15:50 高須夫悟（奈良女子大）空間個体群動態の数理：個体ベースの出生・死亡・空間移動を数理的に記述する方法について

15:50-16:30 阪上雅昭（京都大）変分オートエンコーダによる乳幼児の表出語彙数発達の解析

16:30-17:00 討論その他