生物多様性と群集動態：定量化の数理と統計的推定法

11月8日
14:00-14:30 島谷健一郎（統数研） 生物多様性の統計数理の曼陀羅
14:30-14:40 開催地からのあいさつ＋事務連絡 田中健太（筑波大菅平）
14:40-15:50 深谷肇一（環境研） 長期・広域の生態系モニタリングのモデル化：種多様性の変化を推測する
16:00-17:30 長田穣（東北大） 複雑に相互作用する生物群集と平均場近似

11月9日
9:00-10:30 東樹宏樹（京都大） 大規模データで群集集合の共通原理を探る
10:40-12:10 川津一隆（東北大） ランダム行列の観点から解き明かす生態系の創発特性
13:00-14:30 エクスカーション　菅平高原
15:00-15:50 田中健太（筑波大） 草原の継続期間による植物・昆虫群集の遷移：多様性と種特性の変化
15:50-16:40 門脇浩明（京都大） 理論・統計・シミュレーションの三位一体～あたらしい生態学教育をめざして～
16:40-17:10 島谷健一郎（統数研） 生物多様性指数の統計数理 サンプル被覆度とデータからの推定法

11月10日
9：00－10：30 近藤倫生（東北大） 種間相互作用とは何か：生物学的レベルの重要性
10:30-11:10 島谷健一郎（統数研） 生物多様性指数の統計数理
11:10 – 12:00 総合討論

生物多様性の統計数理の曼陀羅　島谷健一郎

いきなり統計数理の話をします（事務連絡は後回し）

0. ほんの少しの前置き
1. 群集の生物多様性指数：ヒル数
2. 観察されなかった種数に関するChaoの推定量
3. 統計学は学習しずらい…? 群集生態学も…??
4. 曼荼羅のすすめ

明日午後：
・サンプル被覆度 (coverage)
・データからヒル数の推定法

明後日：
・Chao以外の観察されなかった種数推定法
・ヒル数以外の多様性指数とその応用例

はるか昔から変わらぬ日本の科学界全般に共通する問題

・研究者が議論する場が足りない

ずっとずっと、様々な対策は試され続けられてきた

例：新しい学会の立ち上げ

80年代～　種生物学会、動物行動学会、数理生物学会、
生態学会関東地区・修士論文発表会

2000年代～バイオロギング研究会

会費不要の学会：定量生物の会、次世代シーケンサー現場の会

・統計思考院人材育成事業ワークショップも対策の一つ

生物多様性とは？

定性的：様々な生物が互いに影響し合いながら共存している

レベル：種内の遺伝的多様性、群集、ランドスケープ、生態系、…

群集（同じtrophic level, e.g. 森林の樹木群集）, food web, 寄生，共生、…

定量的：生物多様性の減少や回復を論ずるには多様性を「数値で表す」が不可欠だが、**著しく未完成** (Gaston 1996)

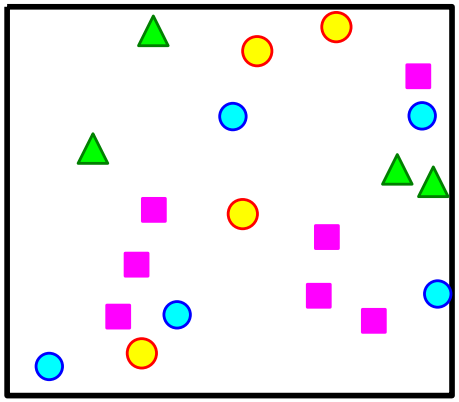
（群集で）**最もよく使われる数値：種数（species richness）**

種の個体数（species abundance）情報もあるとき、

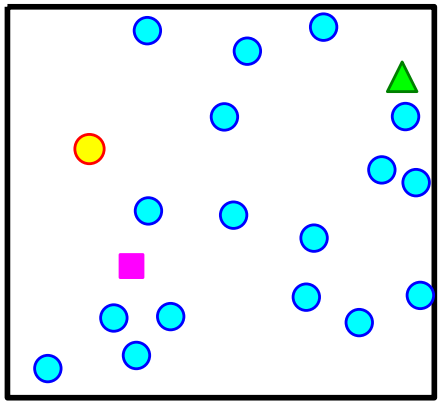
同じ種数でも，

以下、島谷からは群集の種多様性に絞った話題を提供します

どの種も同じくらい



ひとつの種が多い



均一さが違うと，多様性も違うはず

種 s の相対頻度：p_s ランダムに選んだ1個体が種 s である確率

均一さを考慮した指数

- ジニ・シンプソン指数 $D = 1 - \Sigma_s p_s^2$

ランダムに選んだ2個体の種が異なる確率

- シャノン指数 $H = -\Sigma_s p_s \log p_s$

情報量や統計物理のエントロピー

N個の玉をn個の袋にN_1個、N_2個、…、N_n個と入れる時の入れ方の総数の対数 $\ln \frac{N!}{N_1!N_2!\cdots N_n!}$ の$N\to\infty$での極限と考えられる

どれが「最も優れた」生物多様性の数値化か？ **愚問**

- 優占種たちのバランス重視：ジニ・シンプソン指数
- 稀少種の存在を重視：種数

多様性を数値化する目的が何か，どの側面を重視すべきか，よく考えて選ぶ

ところが，でも，しかし，

多様性の数値化として満たすべき性質がある Jost (2006)

多様性は加えられる　replication principle

群集	種構成	ジニ・シンプソン指数	シャノン指数	種数
A	△ △ △ ○ ○	0.48	0.67	2
B	× × × ＊ ＊	0.48	0.67	2
A+B	△ △ △ ○ ○ × × × ＊ ＊	0.74 **単純和にならない**	1.37	4 **単純和**

多様性の数値化として満たすべき性質

1. 共通種を持たない群集を合わせたら単純和
2. 均一のとき ($p_s = 1/S$) 最大, ＝種数 S (不均一のとき $< S$)
3. 既存の指数(の変換)を含む(統一する)

ヒル数 Jost (2006), Hill (1973)

$$ {}^{q}D = \left(\sum_{s=1}^{S} {p_s}^{q}\right)^{\frac{1}{1-q}} \qquad (q \geq 0) $$

有効な種数 という概念

$q = 0$ のとき ${}^{0}D$ は種数 S

$q = 2$ のとき ジニ･シンプソン指数 (の逆数)

$$ {}^{2}D = \frac{1}{\sum_{s=1}^{S} {p_s}^{2}} $$

$q = 1$ のとき シャノン指数(を e の肩)

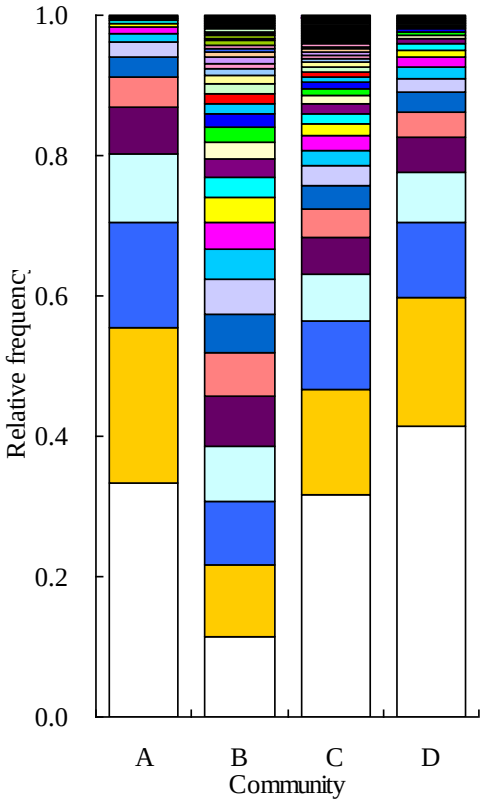
$$ {}^{1}D = e^{-\sum_{s} p_s \ln p_s} $$

ヒル数は共通種を持たないなら単純和

群集	種構成	ジニ・シンプソン指数	シャノン指数	ヒル数 q=2	ヒル数 q=1	ヒル数 q=0 種数
A	△ △ △ ○ ○	0.48	0.67	1.92	1.96	2
B	× × × ＊ ＊	0.48	0.67	1.92	1.96	2
		単純和にならない		単純和		単純和
A+B	△ △ △ ○ ○ × × × ＊ ＊	0.74	1.37	3.85	3.92	4

複数の群集を比較するとき: 1個の数値でなく曲線で比べる

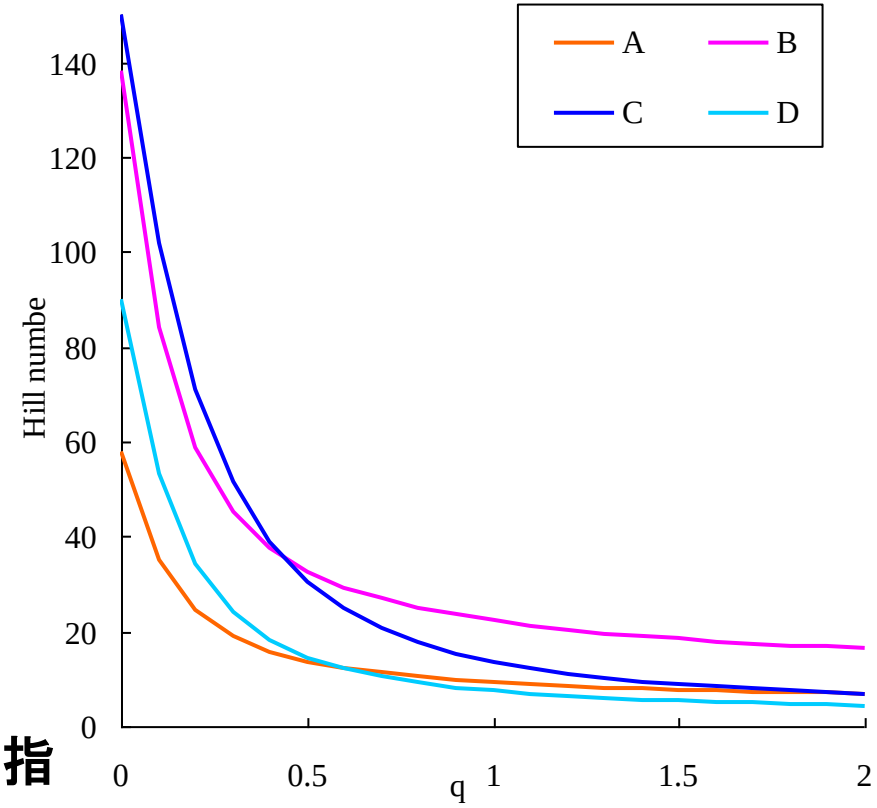
群集の例



いろいろな指数を計算して比べるより

	Species richness	Shannon entropy	Simpson index
A	58	2.23	0.855
B	138	3.10	0.939
C	150	2.61	0.854
D	90	2.01	0.772

ヒル数の曲線を描いて考える



ヒル数(の復権?)により群集の生物多様性指数の問題は決着ついた…ように見えた…

しかし、ヒル数にも重大な欠陥がありました。
ヒル数の数式は群集の種個体数（相対頻度）がわかっている場合のもので、現実には当然、データからの推定値を用います。

N_s: 種 s の観察された個体数 s

$N = \sum_s N_s$: 観察された個体の総数

$\hat{p}_s = N_s / N$:相対頻度の観察値（の最尤推定量）

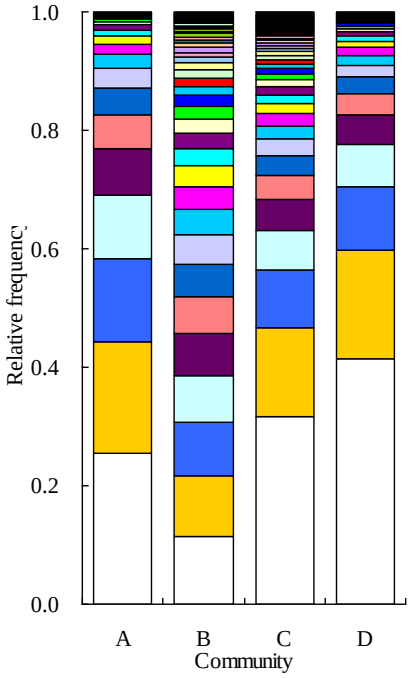
${}^q\hat{D} = (\sum_{s=1}^{S} \hat{p}_s{}^q)^{\frac{1}{1-q}}$:ヒル数の推定量 (plug-in estimator)

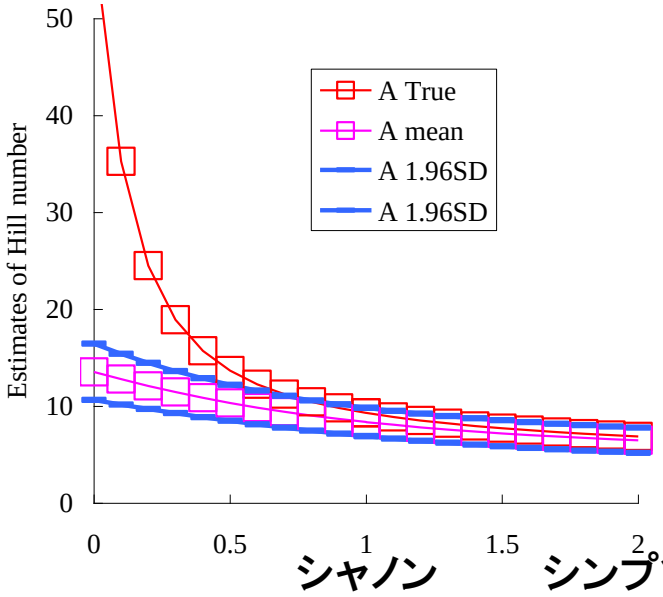
これは真値 ${}^qD = (\sum_{s=1}^{S} p_s{}^q)^{\frac{1}{1-q}}$ を過少推定する

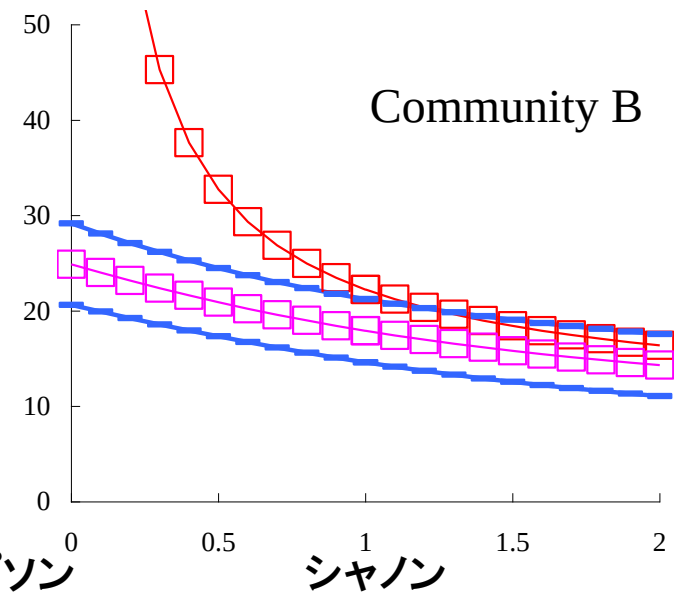
(e.g. Dauby and Hardy 2012)

種 s の在・不在でデータをとった場合も、以下と類似の方法でヒル数を推定できます。なお、背景に来る数学は、多項分布でなくベルヌーイ分布になり、細々と異なります。

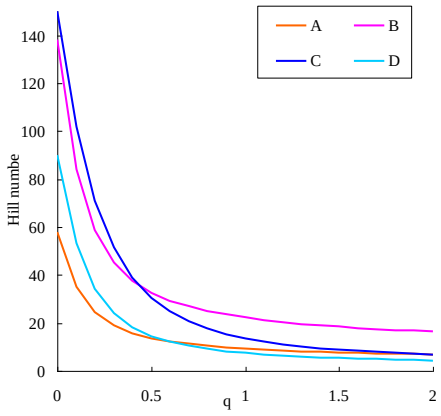
ヒル数の式に観察された相対頻度を代入すると過少推定になる シミュレーションによる例示(サンプル数 $N = 100$)

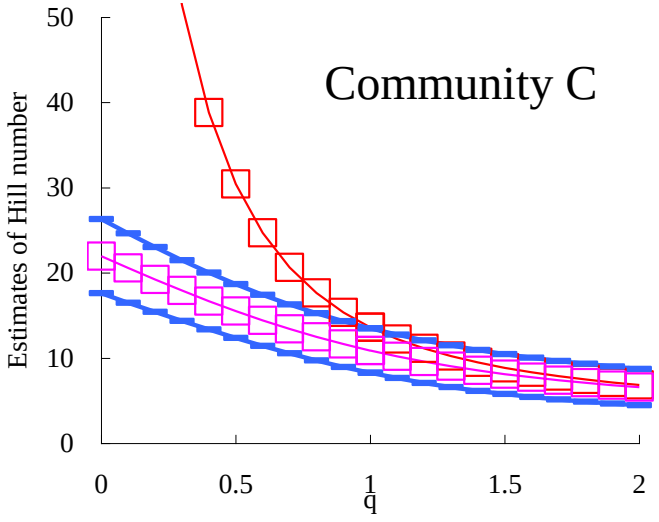


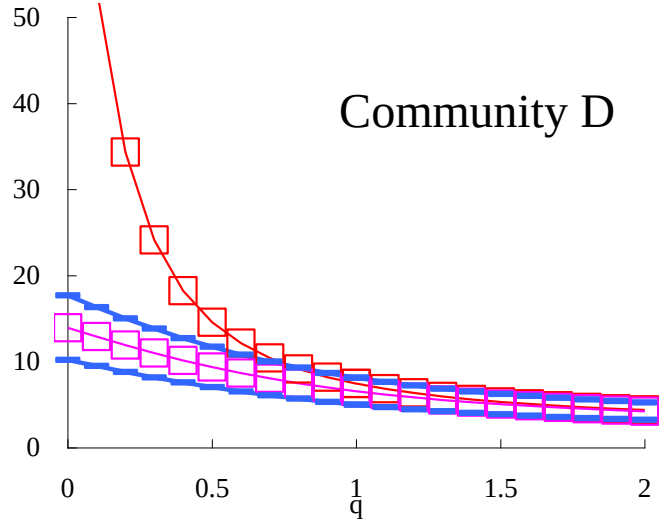




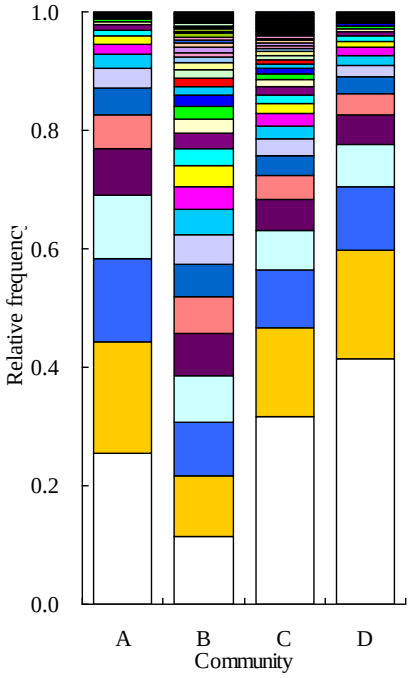
シャノン　　シンプソン　　シャノン

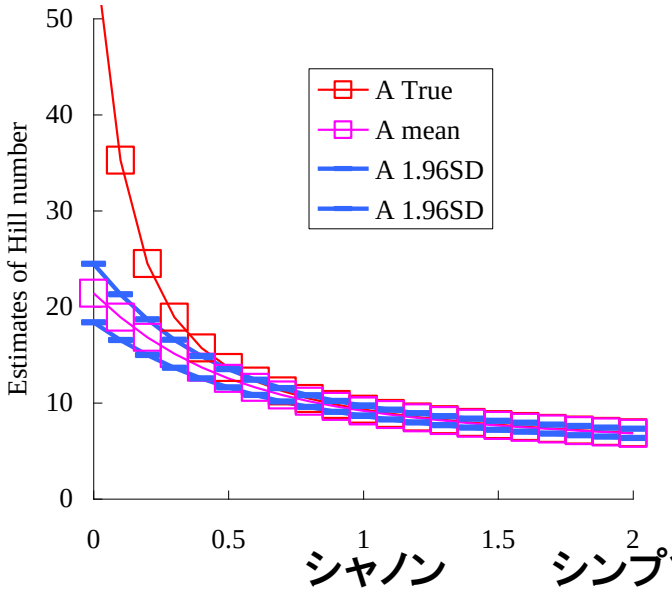


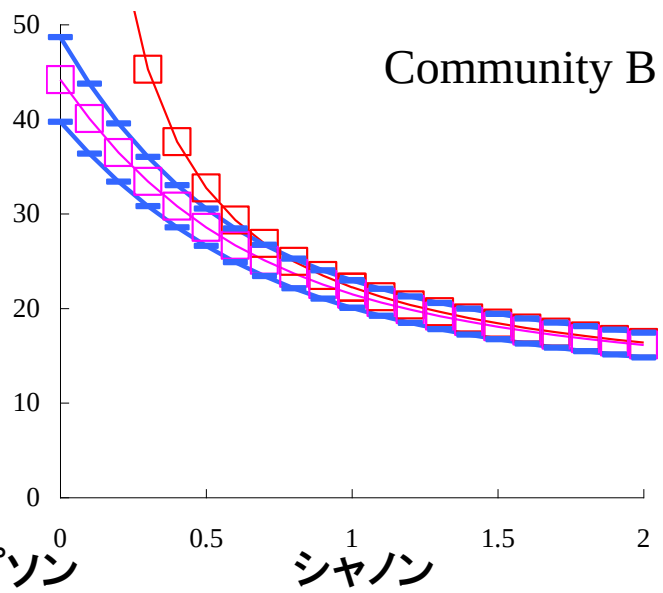


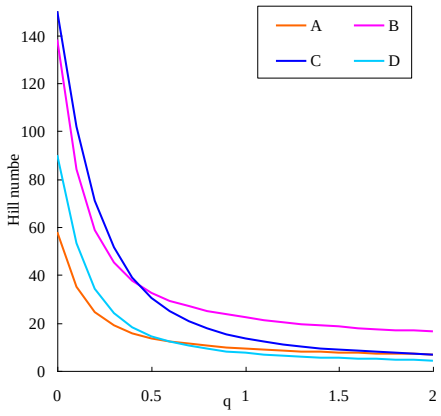


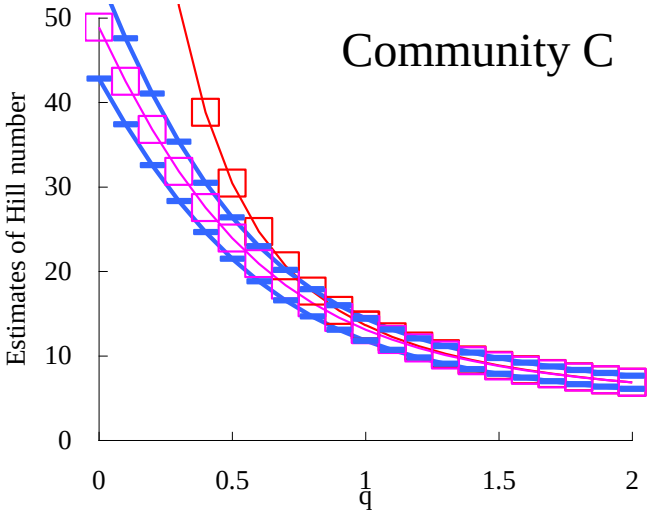
サンプル数をいくら増やしても種数は過少推定になる。でも、観察できなかった種について、補正・推定は不可能…???

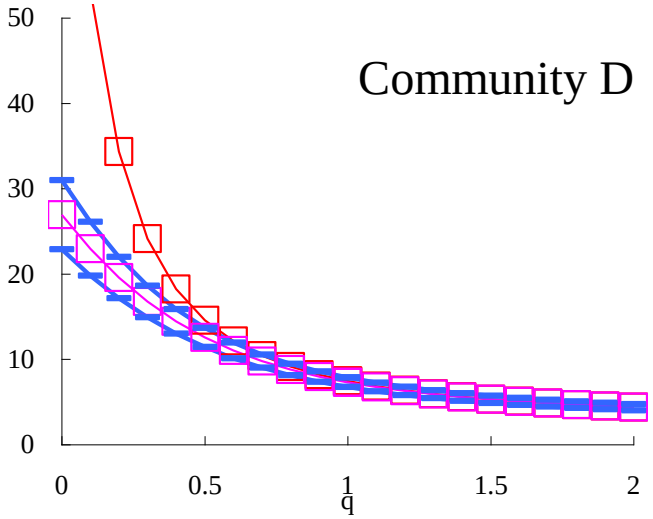












観察されなかった**種の推定は無理**（アタリマエ）

しかし，仮定を置くことで

観察されなかった**種数は推定可能**

ちょっと待って．常識的には

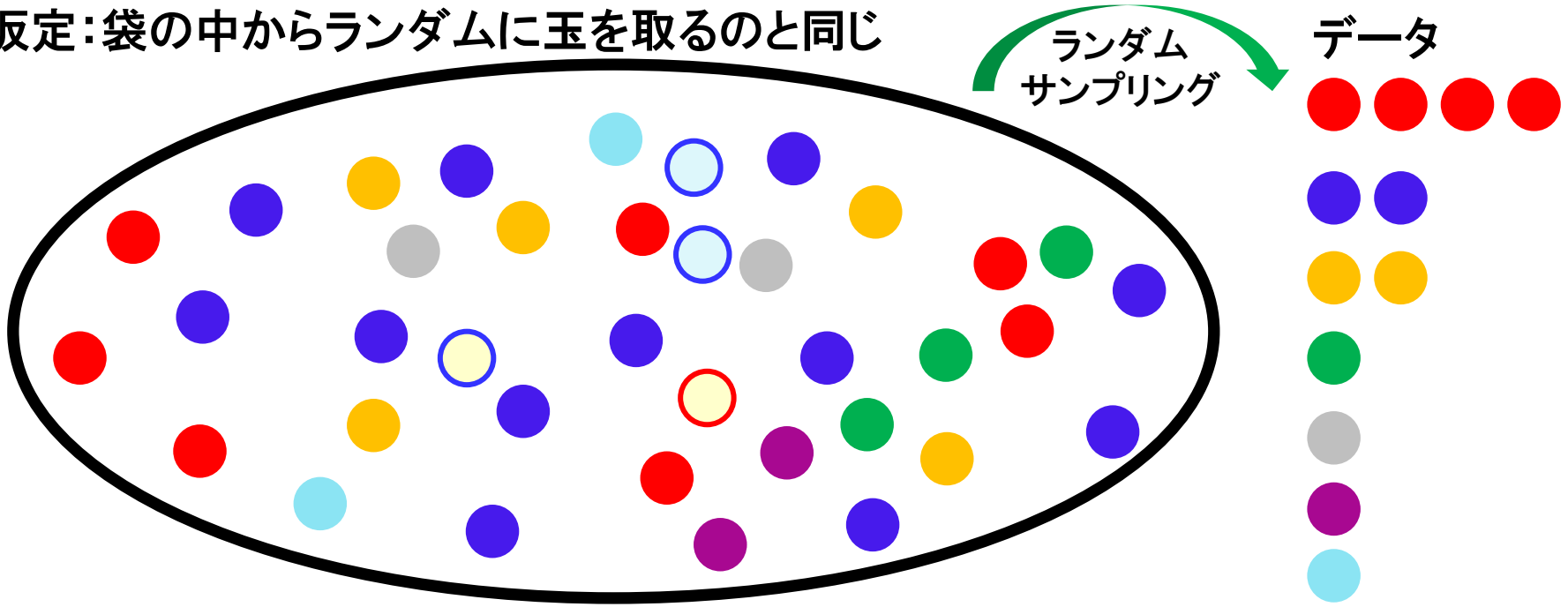
1. 種のリストを作る
2. 種数がわかる

1を経ずにどうして2ができるの？　アホか…

Chaoの推定法

仮定：個体の観察：袋の中からランダムに玉を取るのと同じ確率の仕組みに従う

仮定：袋の中からランダムに玉を取るのと同じ

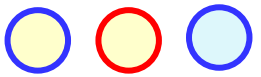


種 s の観察数 X_s ：

サンプルごとに変動する（確率変数）ので大文字 X を使う

データになると確定しているので小文字 x を使う

観察されなかった



10個体観察できた種の数：0

4個体観察できた種の数：1

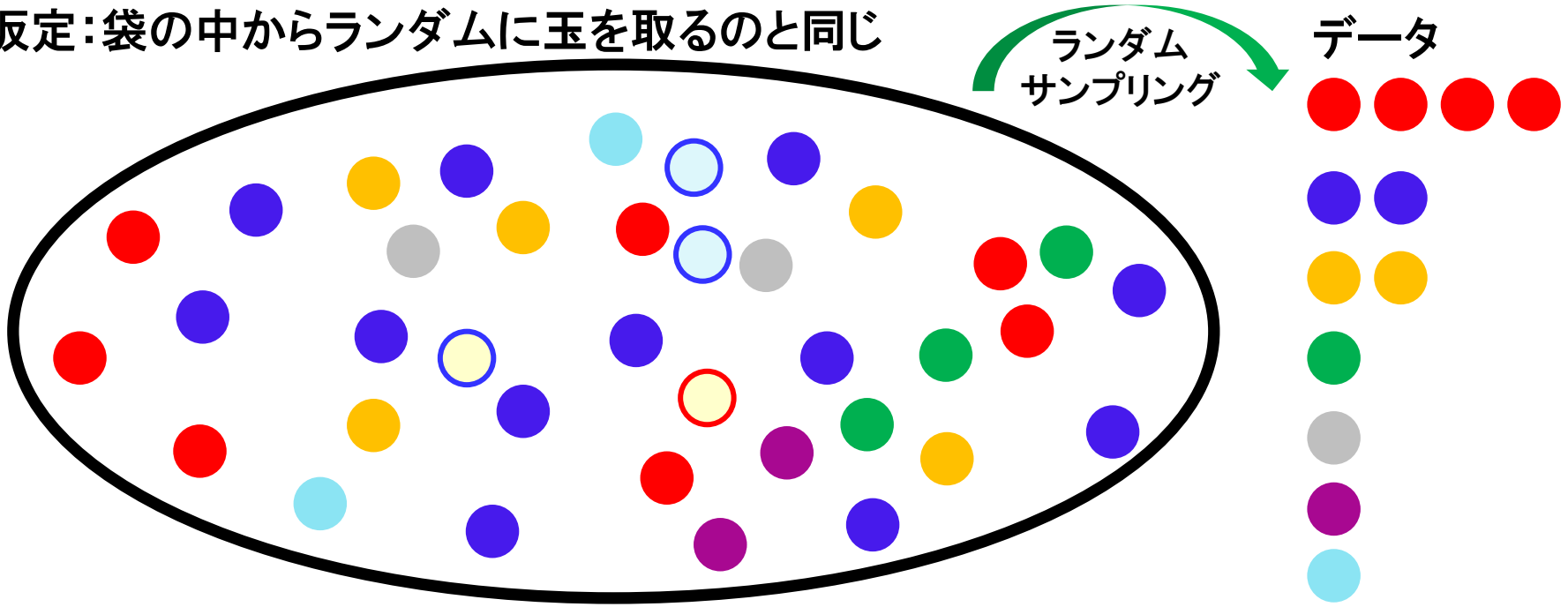
3個体観察できた種の数：0

2個体観察できた種の数：2

1個体観察できた種の数：4

0個体観察できた種の数 = 観察できなかった種数

仮定：袋の中からランダムに玉を取るのと同じ



種 s の観察数 X_s ：
サンプルごとに変動する（確率変数）ので大文字 X を使う
データになると確定しているので小文字 x を使う

観察されなかった

観察された個体数が k である種の数を F_k

4個体観察できた種の数： $f_4 = 1$
3個体観察できた種の数： $f_3 = 0$
2個体観察できた種の数： $f_2 = 2$
1個体観察できた種の数： $f_1 = 4$

F_k は確率変数なので大文字

f_k は観察数（確率変数の実現）なので小文字

0個体観察できた種の数 = 観察できなかった種数: $f_0 = ???$

相対頻度p_sの種 s が k 個体観察される確率は、2項分布を用いて

$$P(X_s = k) = \binom{n}{k} p_s^k (1-p_s)^{n-k} \qquad \binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1} = \frac{n!}{k!(n-k)!}$$

組み合わせ ${}_nC_k$ とも書く

k 個体観察される種の期待値は

$$\mathbf{E}[F_k] = \sum_{s=1}^{S} (P(X_s = k)\cdot 1 + P(X_s \neq k)\cdot 0) = \sum_{s=1}^{S} \binom{n}{k} p_s^k (1-p_s)^{n-k}$$

$P(\ )$: ( ) が起こる確率
E[ ]: [ ]の中の確率変数の期待値

$$\mathbf{E}[F_2] = \frac{n(n-1)}{2} \sum_{s=1}^{S} p_s{}^2 (1-p_s)^{n-2} \qquad \mathbf{E}[F_1] = n\sum_{s=1}^{S} p_s (1-p_s)^{n-1} \qquad \mathbf{E}[F_0] = \sum_{s=1}^{S} (1-p_s)^n$$

種 s の観察数 X_s ：
サンプルごとに変動する(確率変数)ので大文字 X を使う
データになると確定しているので小文字 x を使う

観察された個体数が k である種の数を F_k

4個体観察できた種の数： $f_4 = 1$
3個体観察できた種の数： $f_3 = 0$
2個体観察できた種の数： $f_2 = 2$
1個体観察できた種の数： $f_1 = 4$

F_k は確率変数なので大文字

f_k は観察数(確率変数の実現)なので小文字

0個体観察できた種の数 = 観察できなかった種数: f_0 = ???

相対頻度p_sの種 s が k 個体観察される確率は、2項分布を用いて

$$P(X_s = k) = \binom{n}{k} p_s^k (1-p_s)^{n-k} \qquad \binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1} = \frac{n!}{k!(n-k)!}$$

k 個体観察される種の期待値は

$$\mathbf{E}[F_k] = \sum_{s=1}^{S} (P(X_s = k)\cdot 1 + P(X_s \neq k)\cdot 0) = \sum_{s=1}^{S} \binom{n}{k} p_s^k (1-p_s)^{n-k}$$

$P(\ )$: ( ) が起こる確率
E[ ]: [ ]の中の確率変数の期待値

$$\mathbf{E}[F_2] = \frac{n(n-1)}{2} \sum_{s=1}^{S} p_s^{\,2} (1-p_s)^{n-2} \qquad \mathbf{E}[F_1] = n \sum_{s=1}^{S} p_s (1-p_s)^{n-1} \qquad \mathbf{E}[F_0] = \sum_{s=1}^{S} (1-p_s)^n$$

シュヴァルツの不等式 $\sum_s \alpha_s^{\,2} \sum_s \beta_s^{\,2} \geq (\sum_s \alpha_s \beta_s)^2$

$\alpha_s = (1-p_s)^{\frac{n}{2}}$ $\beta_s = p_s (1-p_s)^{\frac{n}{2}-1}$ を代入

$$\sum_{s=1}^{S} (1-p_s)^n \sum_{s=1}^{S} p_s^{\,2} (1-p_s)^{n-2} \geq (\sum_{s=1}^{S} p_s (1-p_s)^{n-1})^2$$

両辺を $\sum_{s=1}^{S} p_s^{\,2} (1-p_s)^{n-2}$ で割る

$$\sum_{s=1}^{S} (1-p_s)^n \geq \frac{(\sum_{s=1}^{S} p_s (1-p_s)^{n-1})^2}{\sum_{s=1}^{S} p_s^{\,2} (1-p_s)^{n-2}}$$

F_0, F_1, F_2 の期待値について

$$\mathbf{E}[F_0] \geq \frac{\left(\frac{\mathbf{E}[F_1]}{n}\right)^2}{\frac{2\mathbf{E}[F_2]}{n(n-1)}} = \frac{n-1}{n} \cdot \frac{(\mathbf{E}[F_1])^2}{2\mathbf{E}[F_2]}$$

という不等式が得られた

k 個体観察される種の期待値は

$$\mathbf{E}[F_k] = \sum_{s=1}^{S} (P(X_s = k) \cdot 1 + P(X_s \neq k) \cdot 0) = \sum_{s=1}^{S} \binom{n}{k} p_s^k (1-p_s)^{n-k}$$

$P(\ )$: $(\ )$ が起こる確率
$\mathrm{E}[\ ]$: $[\ ]$の中の確率変数の期待値

$$\mathbf{E}[F_2] = \frac{n(n-1)}{2} \sum_{s=1}^{S} p_s^2 (1-p_s)^{n-2} \qquad \mathbf{E}[F_1] = n \sum_{s=1}^{S} p_s (1-p_s)^{n-1} \qquad \mathbf{E}[F_0] = \sum_{s=1}^{S} (1-p_s)^n$$

シュヴァルツの不等式 $\sum_s \alpha_s^2 \sum_s \beta_s^2 \geq (\sum_s \alpha_s \beta_s)^2$

$$\alpha_s = (1-p_s)^{\frac{n}{2}} \qquad \beta_s = p_s (1-p_s)^{\frac{n}{2}-1}$$

を代入

$$\sum_{s=1}^{S} (1-p_s)^n \sum_{s=1}^{S} p_s^2 (1-p_s)^{n-2} \geq (\sum_{s=1}^{S} p_s (1-p_s)^{n-1})^2$$

両辺を $\sum_{s=1}^{S} p_s^2 (1-p_s)^{n-2}$ で割る

$$\sum_{s=1}^{S} (1-p_s)^n \geq \frac{(\sum_{s=1}^{S} p_s (1-p_s)^{n-1})^2}{\sum_{s=1}^{S} p_s^2 (1-p_s)^{n-2}}$$

F_0, F_1, F_2 の期待値について

$$\mathbf{E}[F_0] \geq \frac{\frac{(\mathbf{E}[F_1])^2}{n}}{\frac{2\mathbf{E}[F_2]}{n(n-1)}} = \frac{n-1}{n} \cdot \frac{(\mathbf{E}[F_1])^2}{2\mathbf{E}[F_2]}$$

という不等式が得られた

観察値と期待値は近いだろうから f_1, f_2 で置き換え、f_0 は推定量なので ^ 記号を付けて

$$\hat{f}_0 = \frac{n-1}{n} \cdot \frac{f_1^2}{2f_2}$$

(Chao 1984)

シュヴァルツの不等式

$$\sum_s \alpha_s^2 \sum_s \beta_s^2 \geq (\sum_s \alpha_s \beta_s)^2$$

$$\alpha_s = (1-p_s)^{\frac{n}{2}} \qquad \beta_s = p_s(1-p_s)^{\frac{n}{2}-1}$$

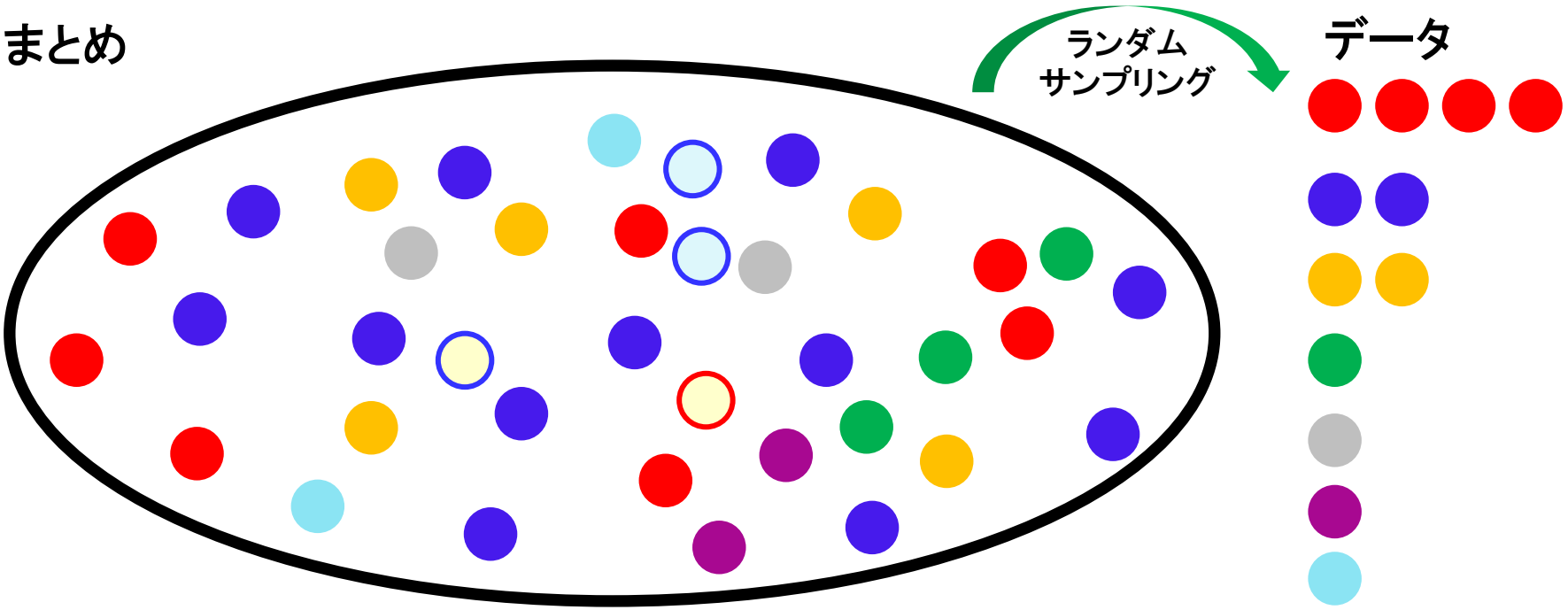
を代入

$$\sum_{s=1}^{S}(1-p_s)^n \sum_{s=1}^{S} p_s^2(1-p_s)^{n-2} \geq (\sum_{s=1}^{S} p_s(1-p_s)^{n-1})^2$$

両辺を $\sum_{s=1}^{S} p_s^2(1-p_s)^{n-2}$ で割る

$$\sum_{s=1}^{S}(1-p_s)^n \geq \frac{(\sum_{s=1}^{S} p_s(1-p_s)^{n-1})^2}{\sum_{s=1}^{S} p_s^2(1-p_s)^{n-2}}$$

まとめ



仮定：袋の中からランダムに玉を取るのと同じ仕組み

観察されなかった

仮定から関係式を作り

観察値(データ)を入れる

k 個体観察できた種の数： f_k

4個体観察できた種の数： f_4

3個体観察できた種の数： f_3

2個体観察できた種の数： f_2

1個体観察できた種の数： f_1

$$\mathbf{E}[F_0] \geq \frac{(\frac{\mathbf{E}[F_1]}{n})^2}{\frac{2\mathbf{E}[F_2]}{n(n-1)}} = \frac{n-1}{n} \cdot \frac{(\mathbf{E}[F_1])^2}{2\mathbf{E}[F_2]}$$

$$\hat{f}_0 = \frac{n-1}{n} \cdot \frac{f_1^2}{2f_2}$$

0個体観察できた種の数 = 観察できなかった種数: f_0

必要な数学 1. 2項分布

$$P(X_s = k) = \binom{n}{k} p_s^k (1-p_s)^{n-k} \qquad \binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1} = \frac{n!}{k!(n-k)!}$$

必要な数学2. 期待値の計算（定義）

$$\mathbf{E}[F_k] = \sum_{s=1}^{S} (P(X_s = k)\cdot 1 + P(X_s \neq k)\cdot 0) = \sum_{s=1}^{S} \binom{n}{k} p_s^k (1-p_s)^{n-k}$$

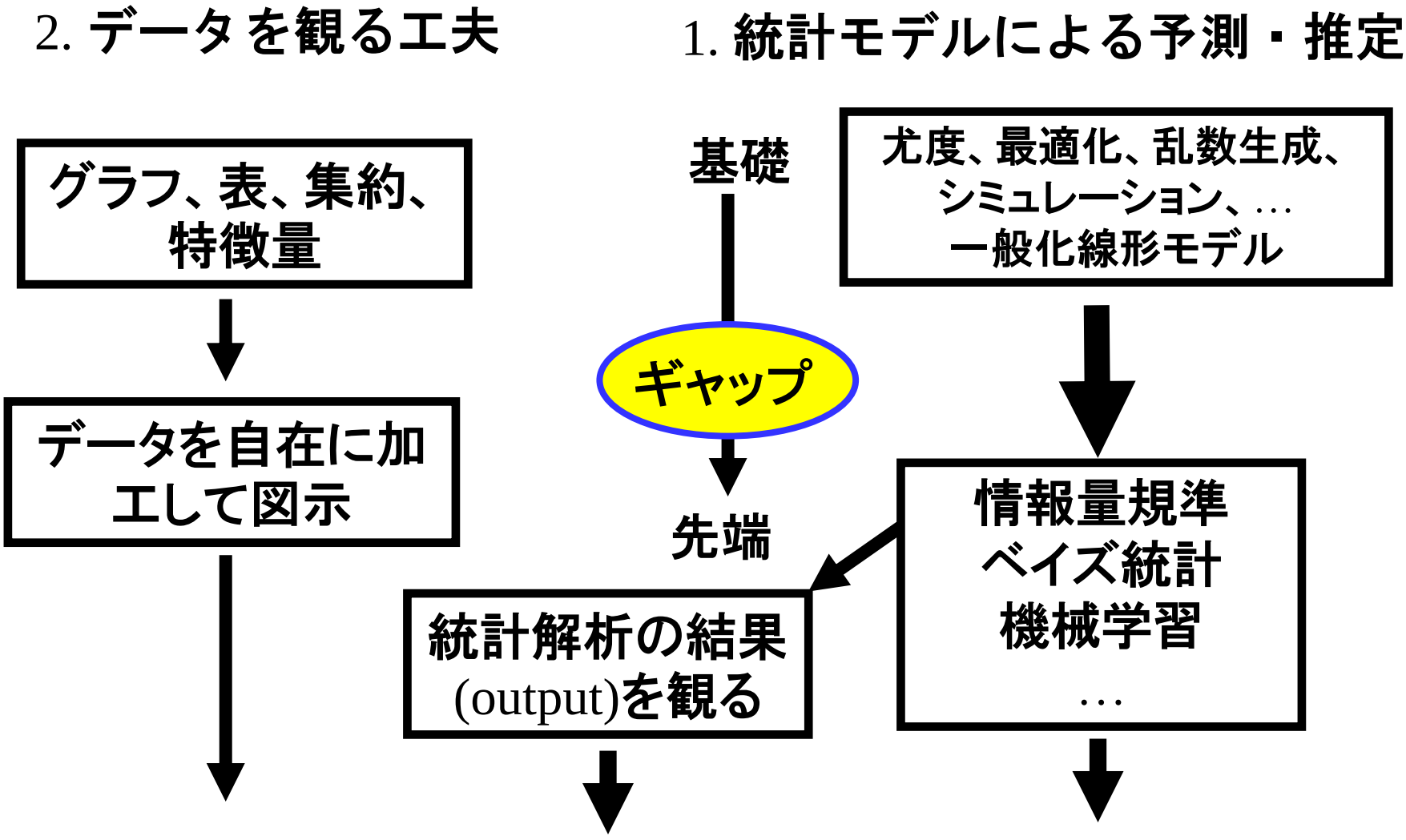
必要な数学3. シュヴァルツの不等式

$$\sum_s \alpha_s{}^2 \sum_s \beta_s{}^2 \geq (\sum_s \alpha_s \beta_s)^2$$

必要な予備知識は持っているはずなのに、聞いてぱっとわかる感じがしない…???

どうして？

統計学（島谷版の全体像）



学部で（つまらない）統計学を履修した．GLMもRで実践した．
なのにどうしてベイズ統計へのハードルは高いままなの？

統計学（島谷版の全体像）

基礎と発展のギャップの例：
観察されなかった種数の推定

2項分布、期待値、
シュヴァルツの不等式

ギャップ??

Chaoの推定量

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

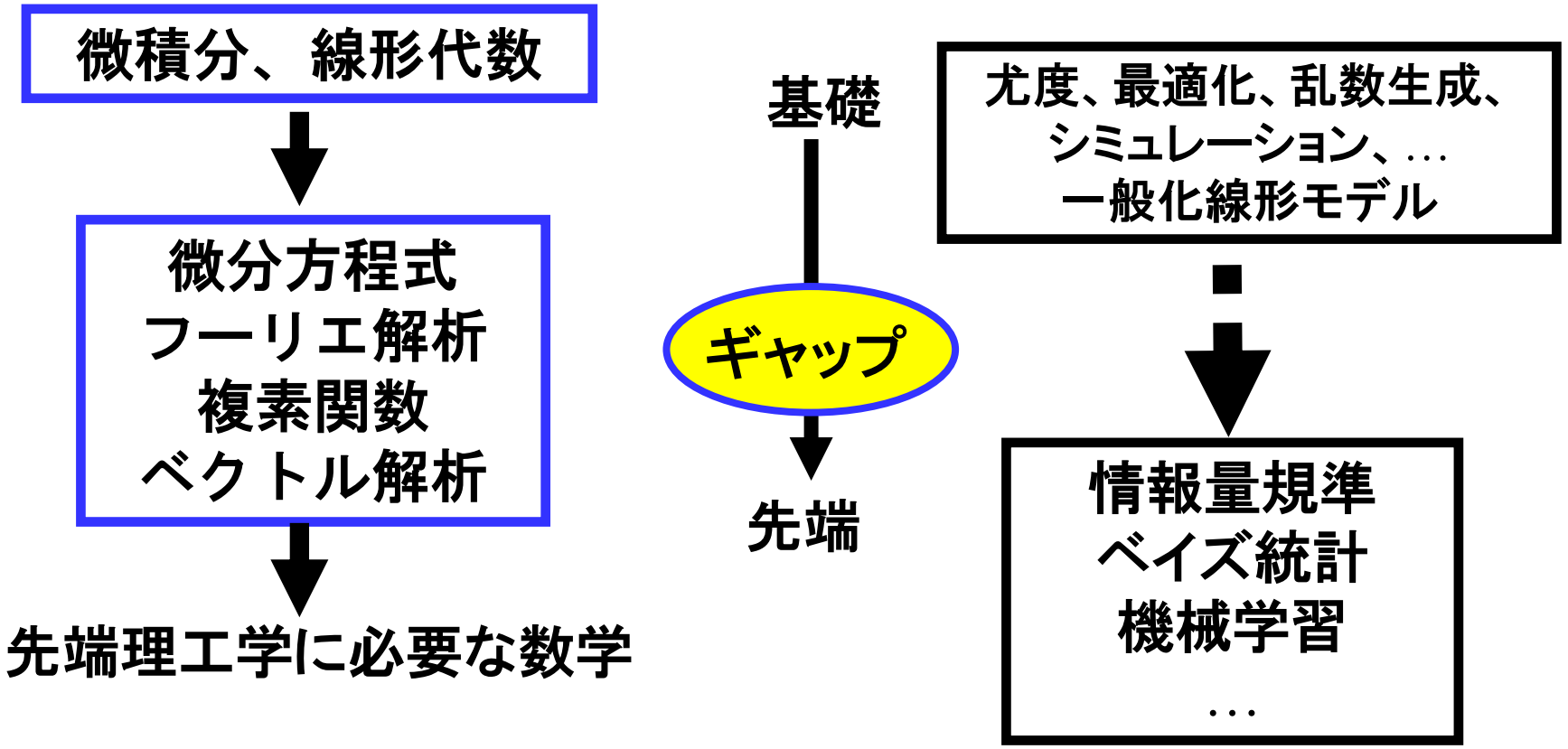
先端

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

学部で（つまらない）統計学を履修した．GLMもRで実践した．
なのにどうしてベイズ統計へのハードルは高いままなの？

理工学の数学（解析学）との比較



＞ 100年

確立してからの年月の差？（島谷説）

＜ 100年

理工学の数学（解析学）との比較

微積分、線形代数

微分方程式
フーリエ解析
複素関数
ベクトル解析

先端理工学に必要な数学

基礎

ギャップ

先端

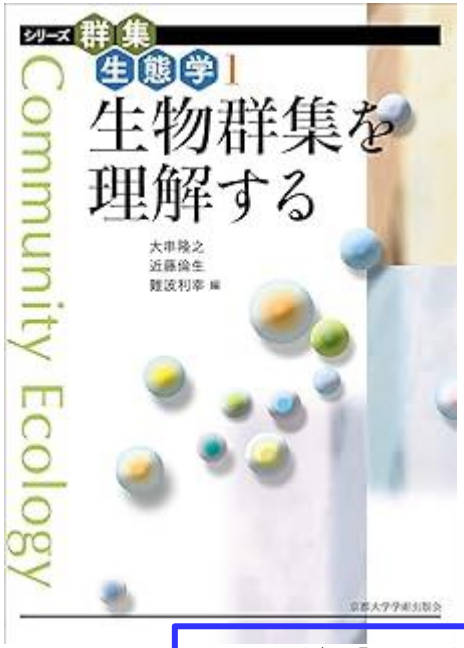
尤度、最適化、乱数生成、
シミュレーション、…
一般化線形モデル

情報量規準
ベイズ統計
機械学習
…

現時点での対処法（島谷提案）

三中信宏氏発案の **曼陀羅**

曼荼羅



フィールドワーク

多様性指数　群集

絶滅　food web　寄生　共生　侵入

環境収容力　ニッチ理論　移入

競争排除則　中立理論

内的成長率　ギルド　採餌

科学哲学　統計モデルによる推定・予測　群集動態、数理モデル

群集生態学に法則はあるか？ (Lawton 1999)

集団生物モデルでは一般性・現実性・正確性にトレードオフが働く (Levins 1966)

群集生態学の現状も似たような感じでないか？

1. 基礎と先端にギャップがある。
2. 各研究者は様々な群集生態学の曼荼羅を彷徨(?)しながら自身を進歩させている…???