ゲノム情報からの
生命現象・病理現象の統計解析

京都大学(医)　統計遺伝学分野
山田　亮
ryamada@genome.med.kyoto-u.ac.jp

今日の内容

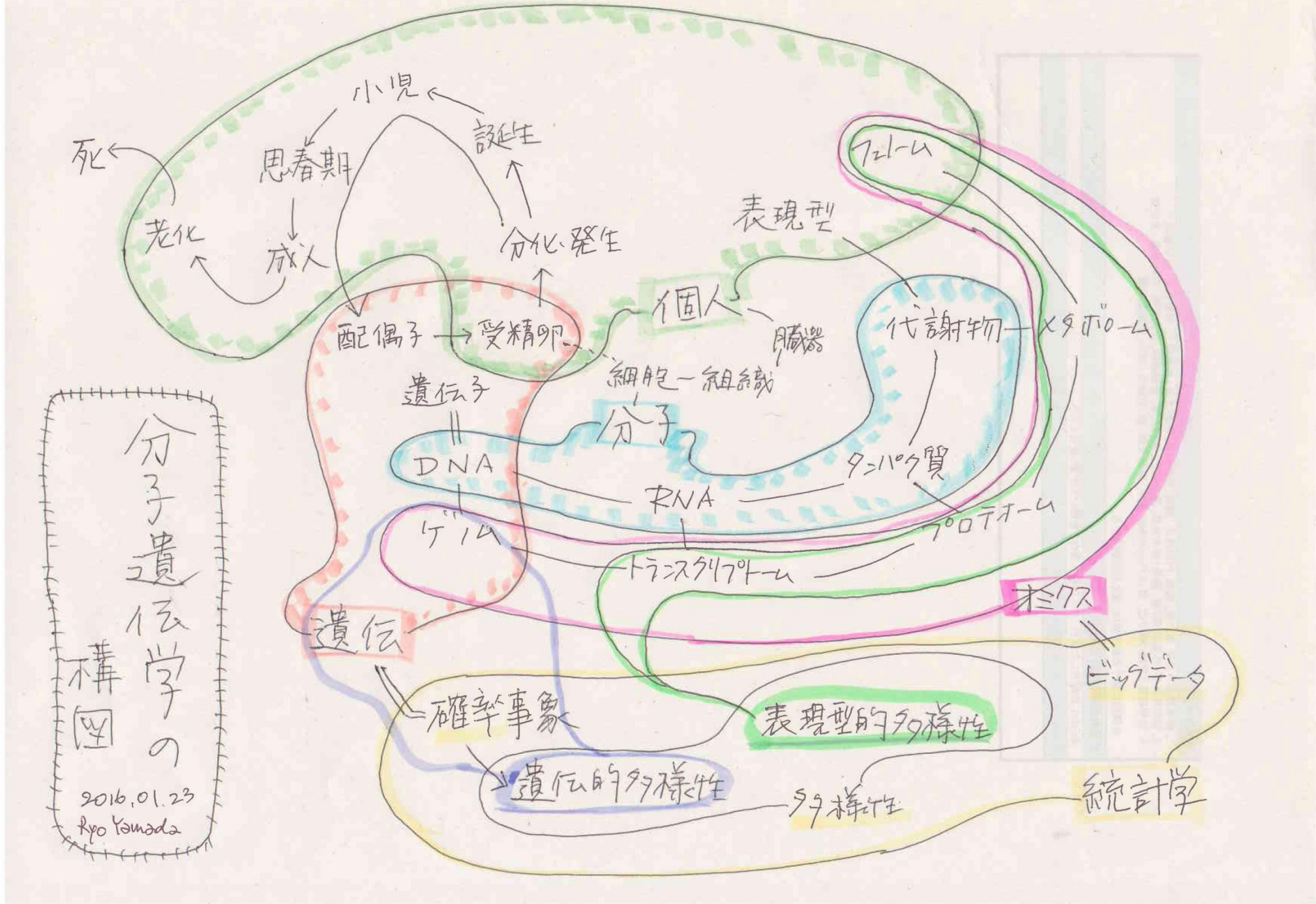
- ジェノタイプとフェノタイプ～解析用にデータを取ること～

- 統計解析手法の俯瞰

今日の目標

- 全体像をつかむ
- 個別のことの概念的な理解をめざし、「細かい理解」は目指さない
- 個別に詳しく知りたいと思ったときのための、「単語のリスト」を入手する
- 取扱い範囲は広いが、それらは色々なところで相互につながりあっていたり、基本的な考え方の組み合わせの諸相だったりすることを理解する

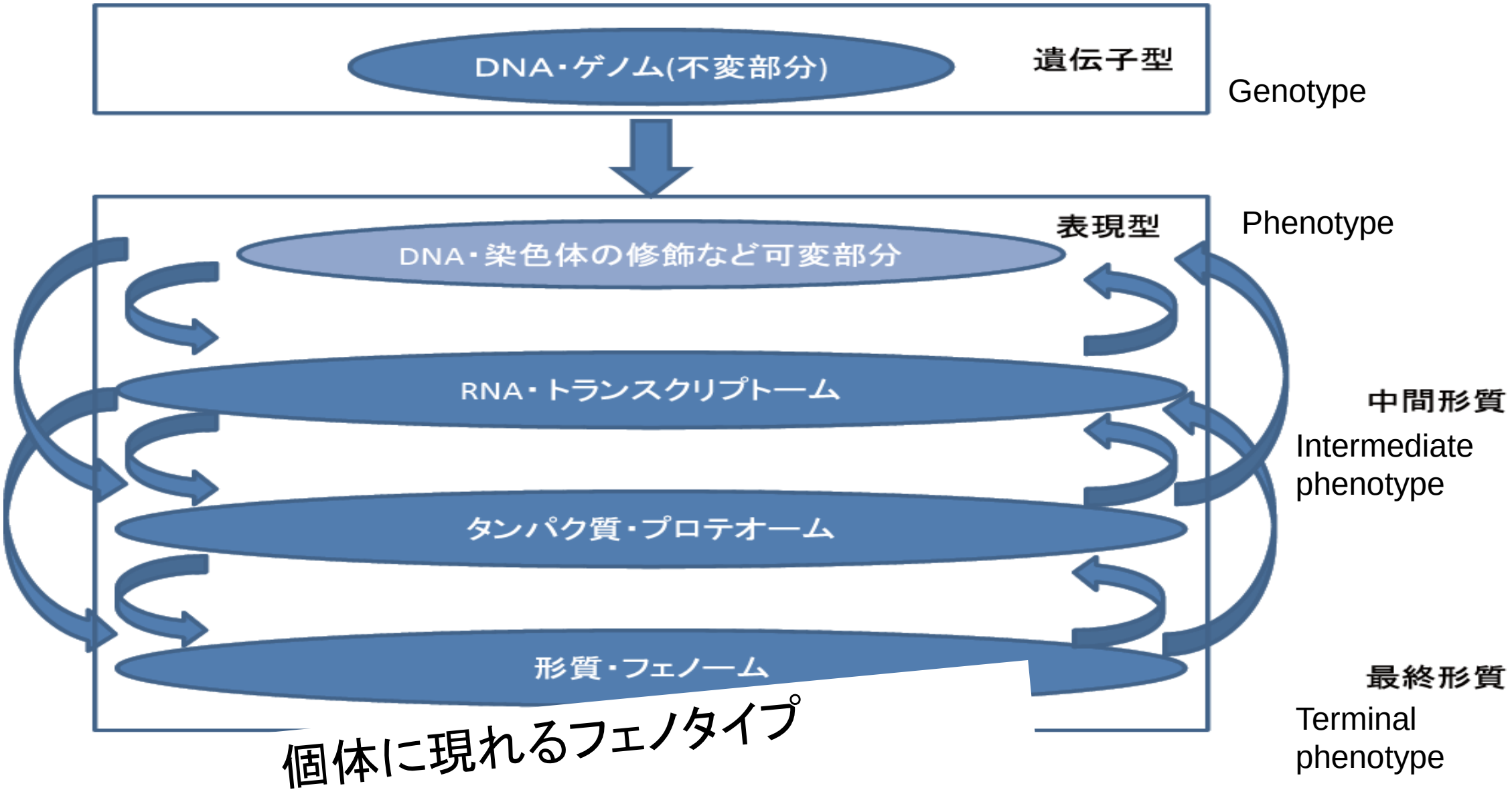
ジェノタイプとフェノタイプ
～解析用にデータを取ること～


分子遺伝学の構図
2016.01.23
Ryo Yamada
小児
誕生
死
思春期
老化
成人
分化・発生
配偶子
受精卵
個人
臓器
細胞ー組織
分子
遺伝子
DNA
RNA
ゲノム
遺伝
表現型
フェノーム
代謝物
メタボローム
タンパク質
プロテオーム
トランスクリプトーム
オミクス
ビッグデータ
確率事象
遺伝的多様性
表現型的多様性
多様性
統計学

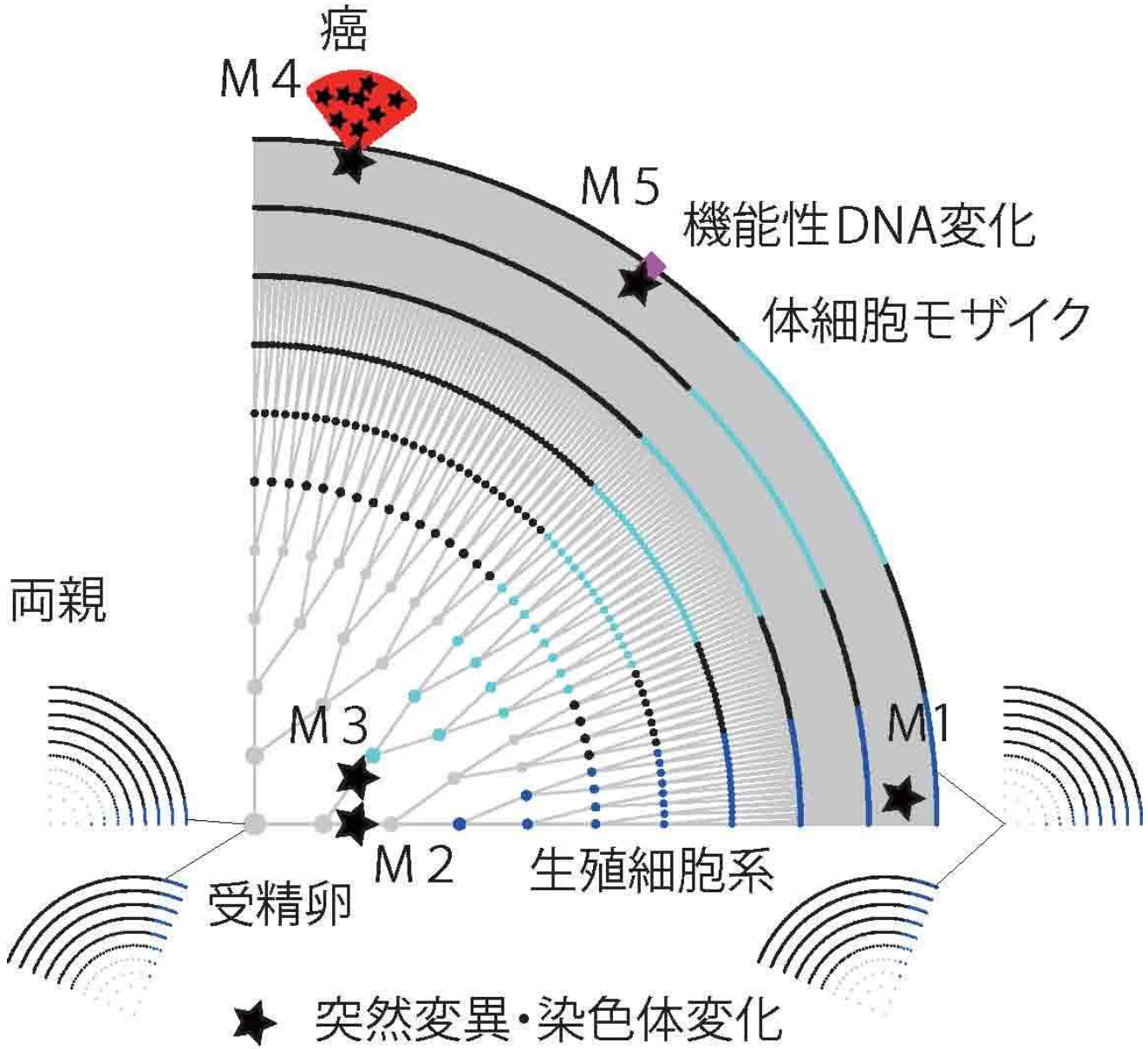
ジェノタイプとフェノタイプ

- 時空間的に一意

- 時空間的に多様


DNA・ゲノム(不変部分)
遺伝子型
Genotype
表現型
Phenotype
DNA・染色体の修飾など可変部分
RNA・トランスクリプトーム
タンパク質・プロテオーム
形質・フェノーム
中間形質
Intermediate phenotype
最終形質
Terminal phenotype
個体に現れるフェノタイプ

個体の時空間

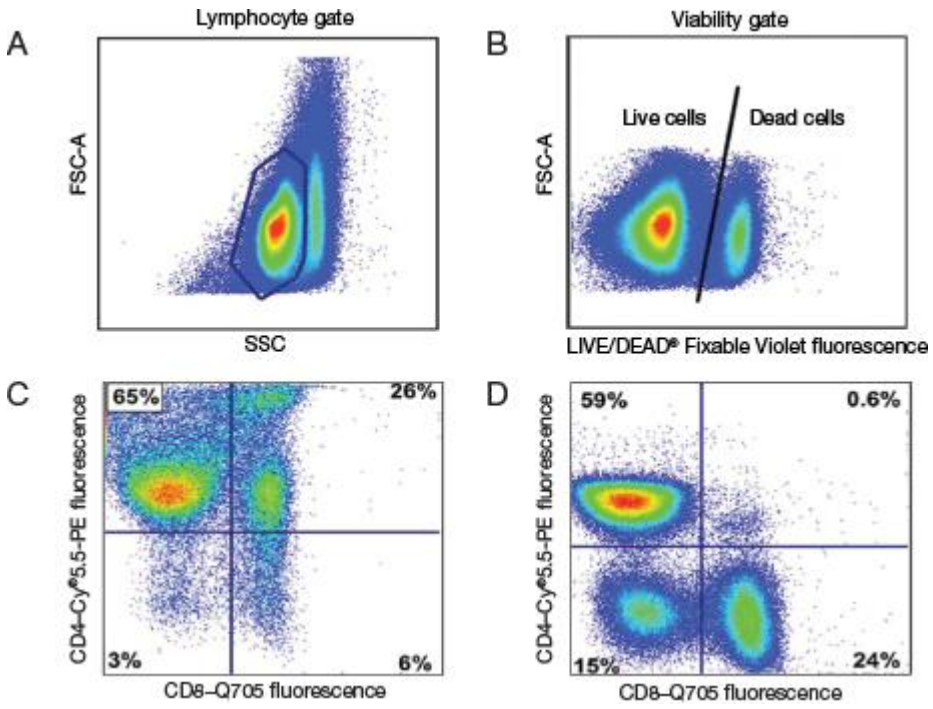


フェノタイプの多様性

- 測定しやすいもの・測定しにくいもの
- 代表値 vs. 分布
- 相互に独立なもの多数 vs. 相互に依存しているもの多数

代表値 vs. 分布

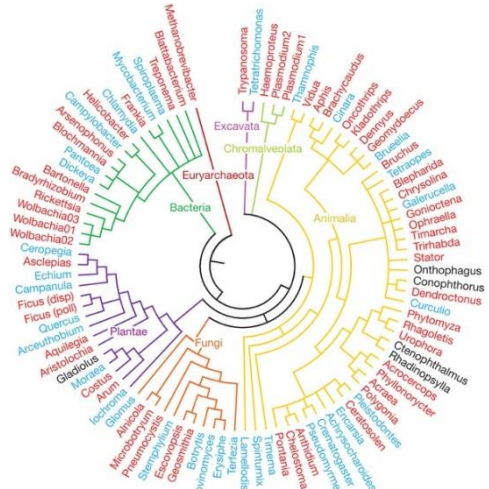
- 温度
  - 気体分子集団の代表値
- 独立試行の多数回測定
  - きれいな分布→代表値→パラ
  - きれいでない分布→分布そのものを→ノン・パラ
- １標本が多観測からなるとき
  - １標本が分布→代表値で大丈夫か？



ThermoFisher Scientific社

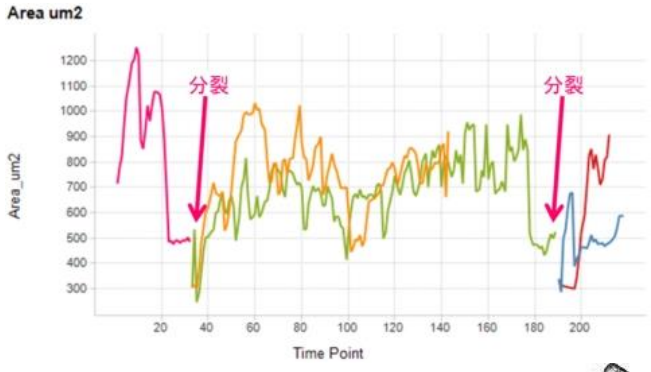
相互に独立なもの多数 vs. 相互に依存しているもの多数

- 相互に相関が強い複数の観測変数

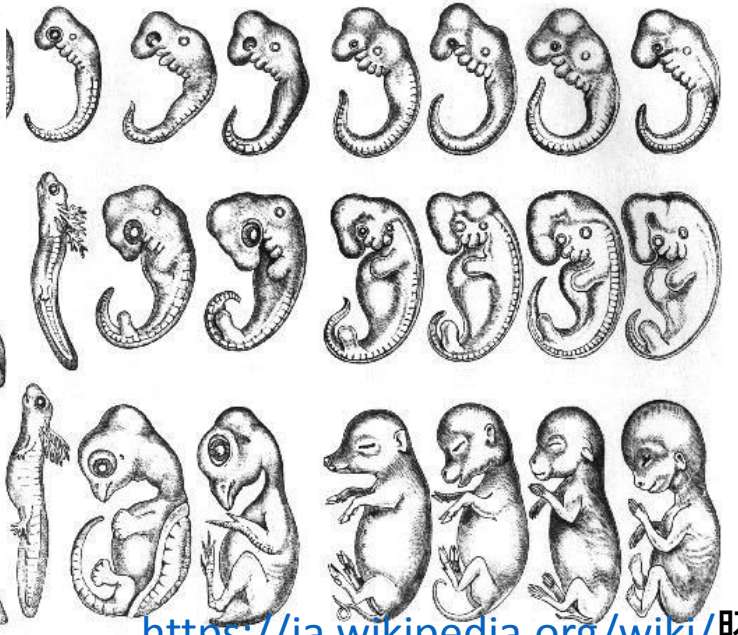


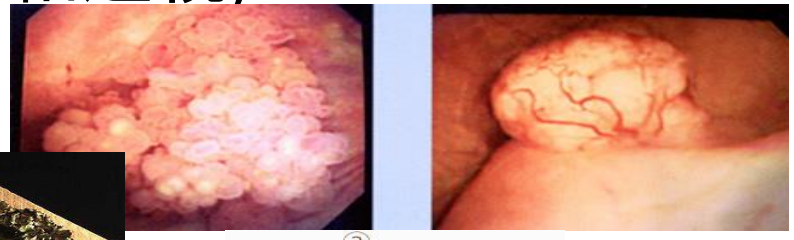
- 時系列データ(時間軸連続)
- 形データ(空間軸連続)
- 運動データ(時空間連続)
- パターンデータ(情報軸連続)

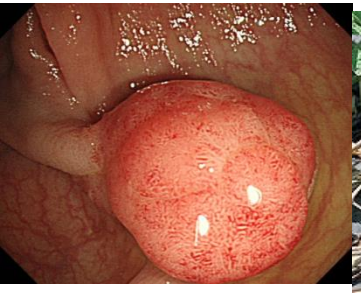
横河電機



Nature 465, 918–921 (17 June 2010)







https://ja.wikipedia.org/wiki/胚

まとめ：ジェノタイプ・フェノタイプという値

- データ解析するために
  - 「値」として取り出す
  - 「値」にも色々
  - いわゆる「値」とは、「数」
  - 「数」とは
    - 自然数・整数・有理数・実数・複素数・ベクトル・行列…
  - いわゆる「値」ではない、データ解析用の「値」とは
    - 数理モデル
      - 特に、生物現象では、ばらつきがあることが基本なので
      - 確率モデル・統計モデル
    - ただし、モデルは(広義の)パラメタで構成するので
    - パラメタの「値」を扱うと言う意味では、「数」に戻る

  - 「いわゆる値」は単純な数理・確率モデルでのパラメタ値
  - より複雑な「タイプ」は複雑なモデルでのパラメタ値

今日の内容

- ジェノタイプとフェノタイプ～解析用にデータを取ること～

- 統計解析手法の俯瞰

統計解析手法の俯瞰

ゲノム・オミクス研究における統計・データサイエンスの役割

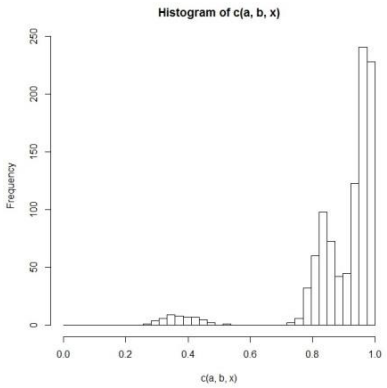
- ノイズのあるハイスループットデータのデータQC
- 検定・推定・分類
- 多次元・高次元データ
- 乱数を使ったアプローチ
- その他：実験デザイン

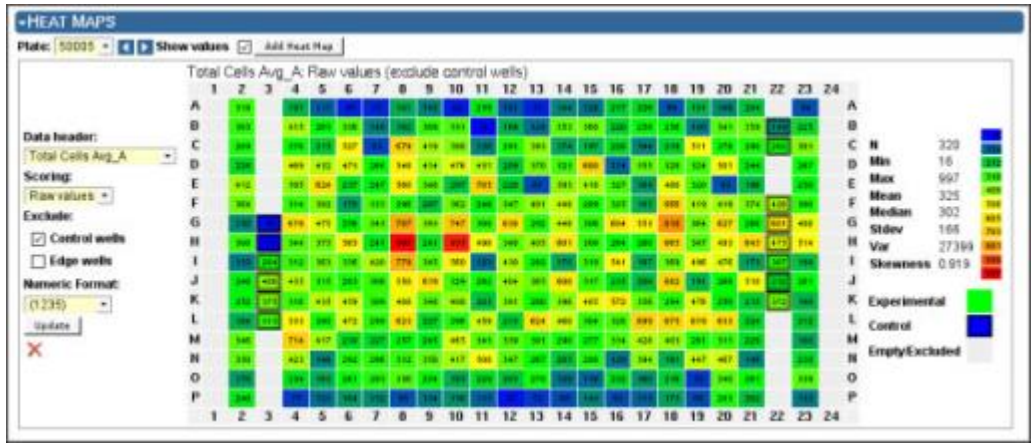
ゲノム・オミクス研究における統計・データサイエンスの役割

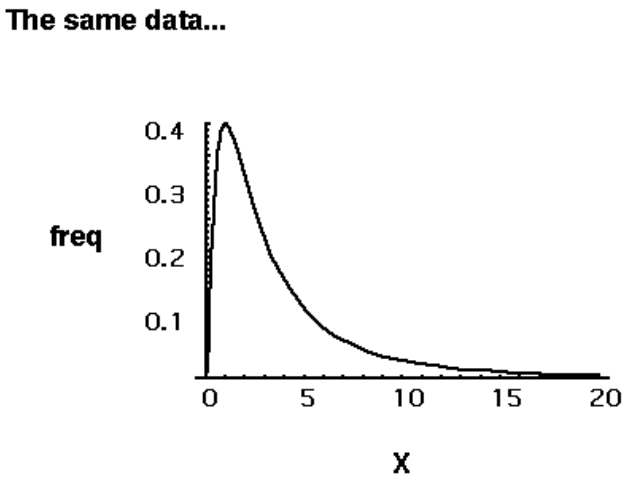
- ノイズのあるハイスループットデータのデータQC
- 検定・推定・分類
- 多次元・高次元データ
- 乱数を使ったアプローチ
- その他：実験デザイン

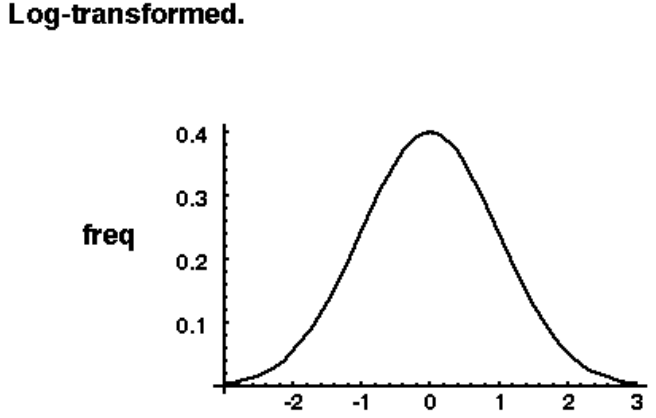
ノイズのあるハイスループットデータのデータQC

- 系統的な誤差/バイアス;サンプル,試薬/実行日/機器/担当者の影響
- ノイズを補正する･コントロールする
  - 外れ値
  - 変換する、1関数で
  - 「場所的効果」について正規化する
  - “コントロール用サンプル”

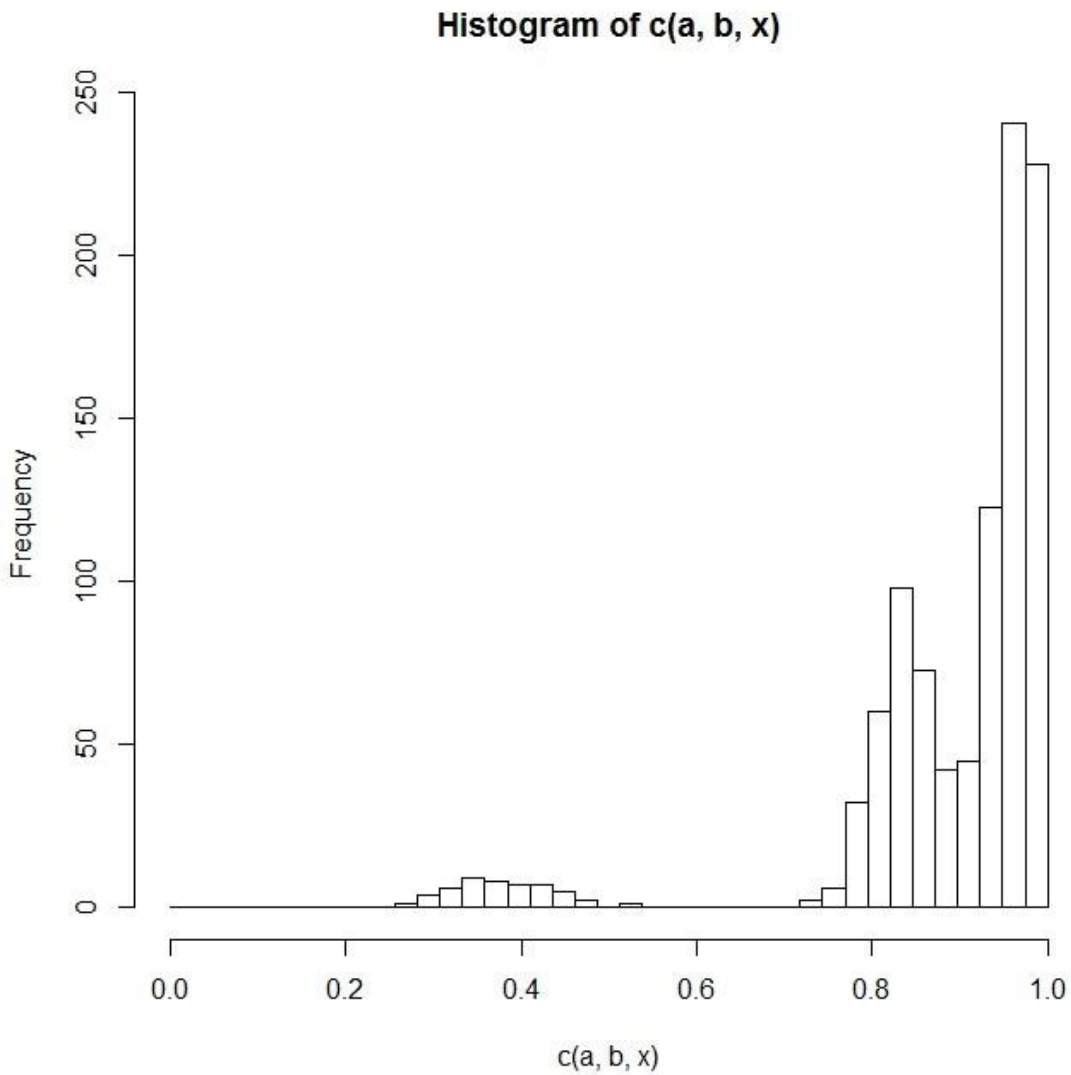




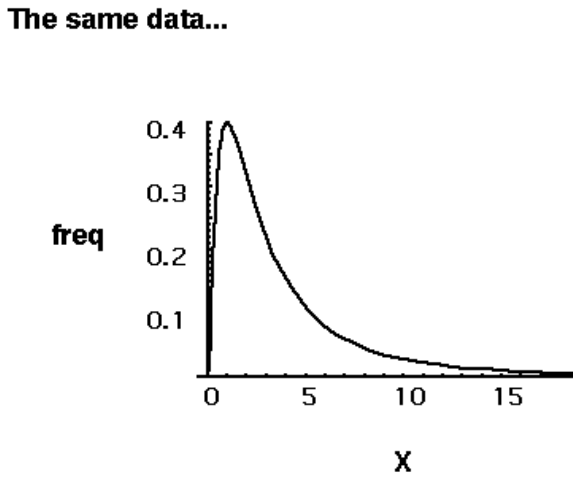


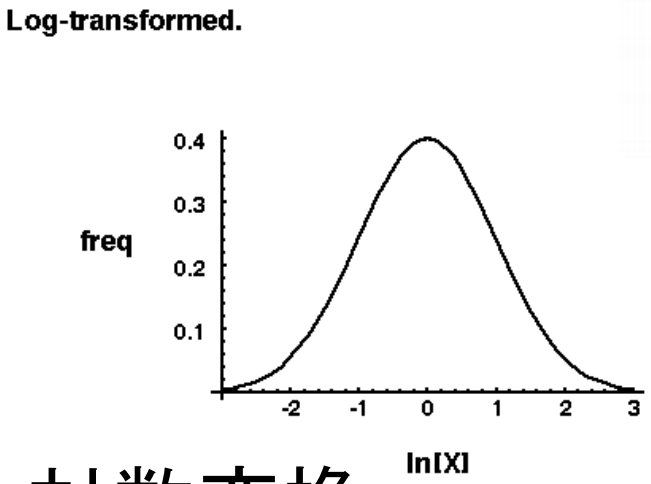


外れ値

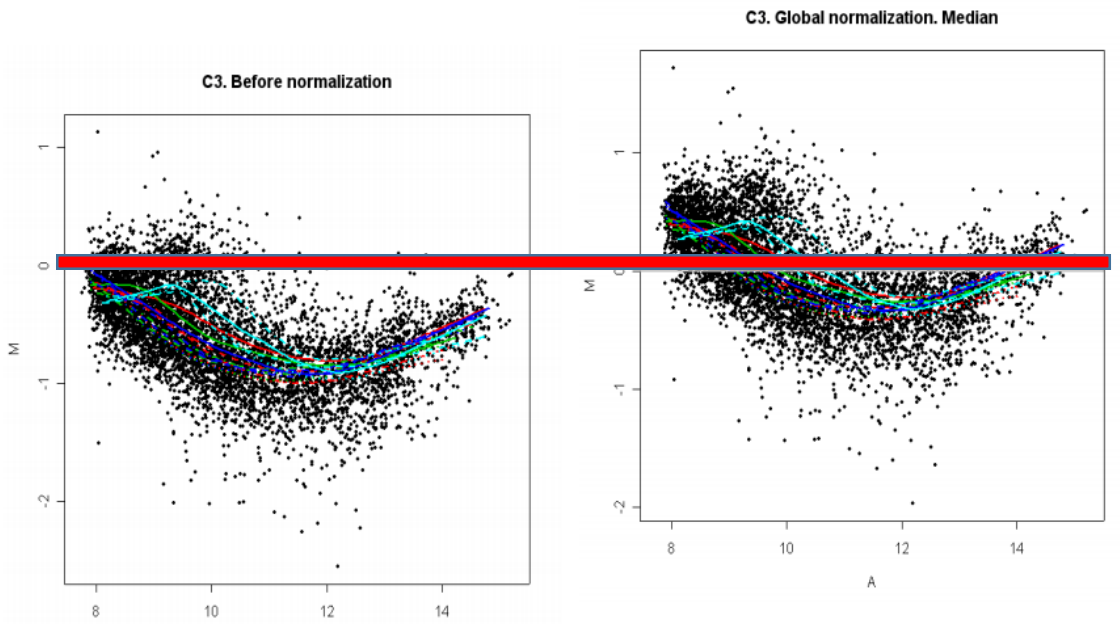

Histogram of c(a, b, x)
Frequency
0
50
100
150
200
250
0.0
0.2
0.4
0.6
0.8
1.0
c(a, b, x)

変換する、1関数で



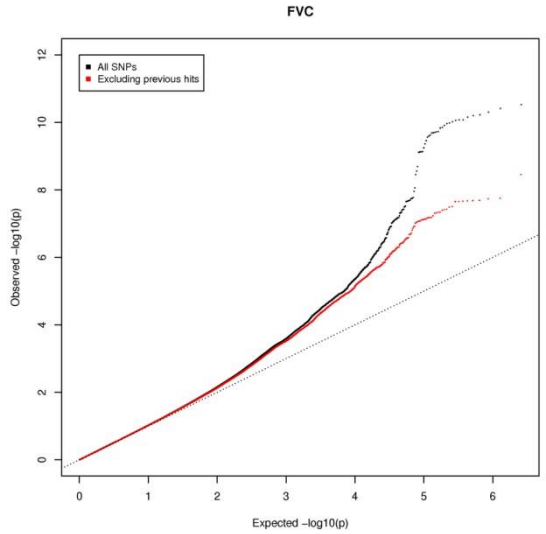


対数変換



中央値を使ったマイクロアレイデータの変換

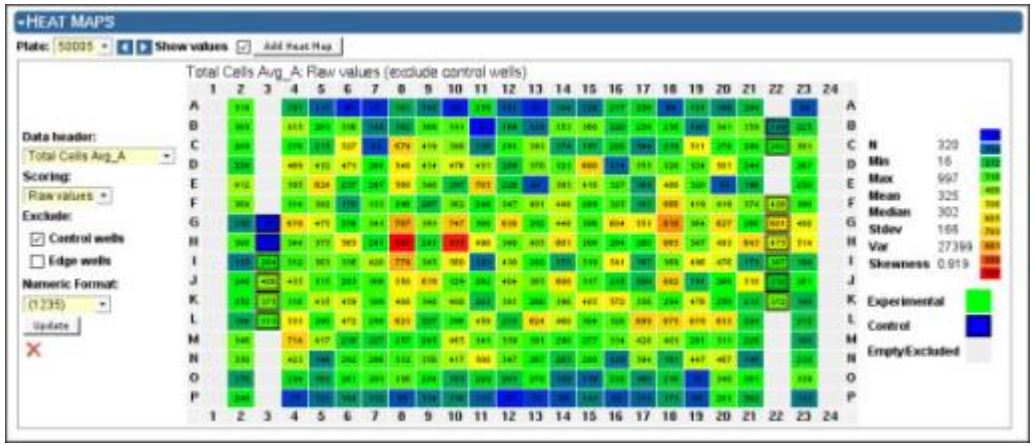
GWASのジェノミックコントロール



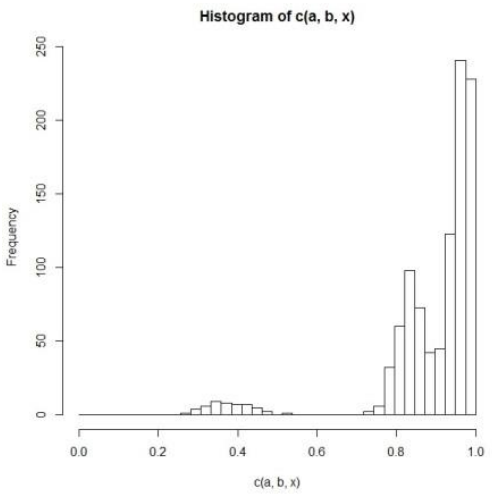
「場所的効果」について正規化する

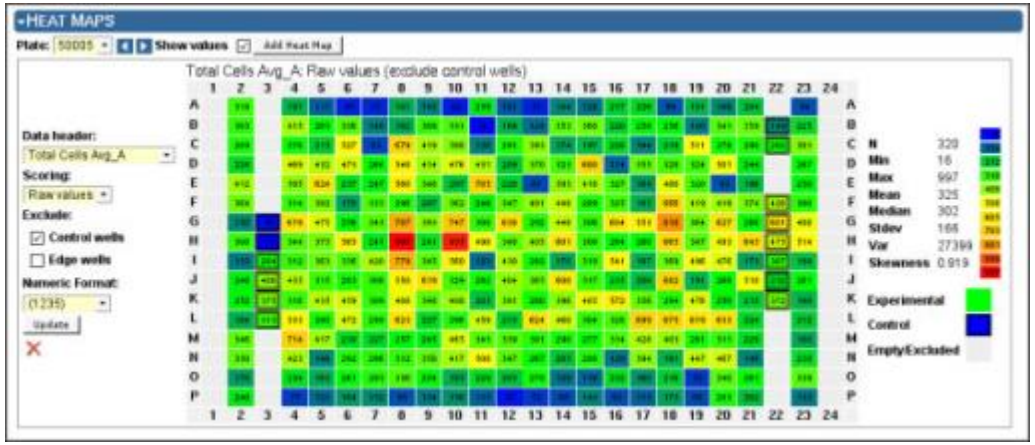
- 「くせ」を取り入れる

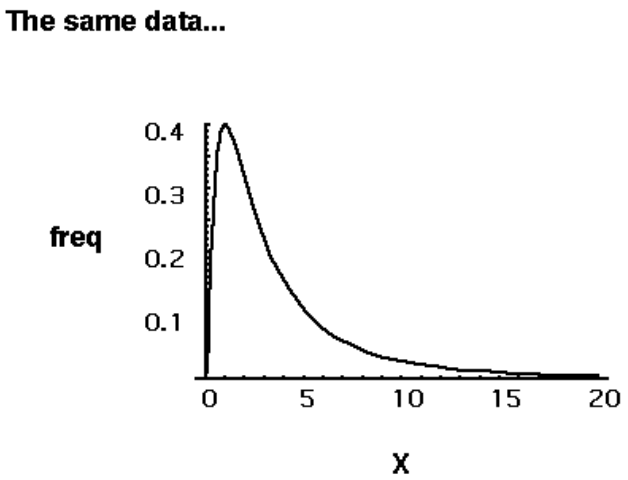
- 単なる「場所」だけでなく、バッチエフェクト全般にも使える
  - データそのものを使わず、事前情報で補正
  - データとその分布を利用して補正

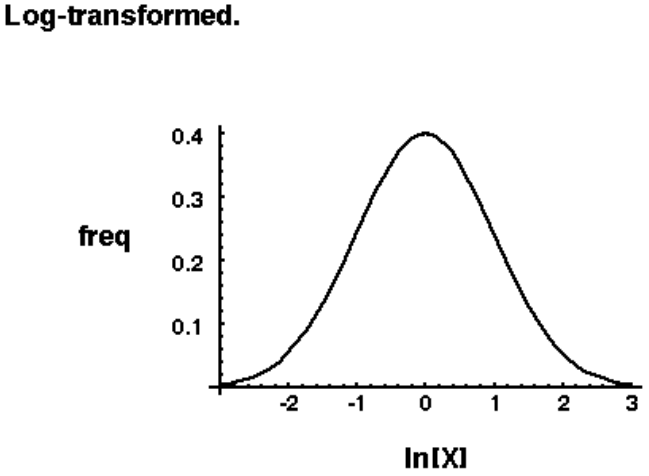


- 系統的な誤差/バイアス;サンプル,試薬/実行日/機器/担当者の影響
- ノイズを補正する・コントロールする
  - 外れ値
  - 変換する、1関数で
  - 「場所的効果」について正規化する
  - “コントロール用サンプル”









ゲノム・オミクス研究における統計・データサイエンスの役割

- ノイズのあるハイスループットデータのデータQC
- 検定・推定・分類
- 多次元・高次元データ
- 乱数を使ったアプローチ
- その他：実験デザイン

検定・推定・分類

- 検定
  - 有意、エラーのコントロール、多重検定
- 推定
  - 区間推定、モデル推定、ベイズ
- 分類
  - 教師アリ、教師ナシ

検定・推定・分類

- 検定
  - 有意、エラーのコントロール、多重検定
- 推定
  - 区間推定、モデル推定、ベイズ
- 分類
  - 教師アリ、教師ナシ

多重検定

p 値とq 値

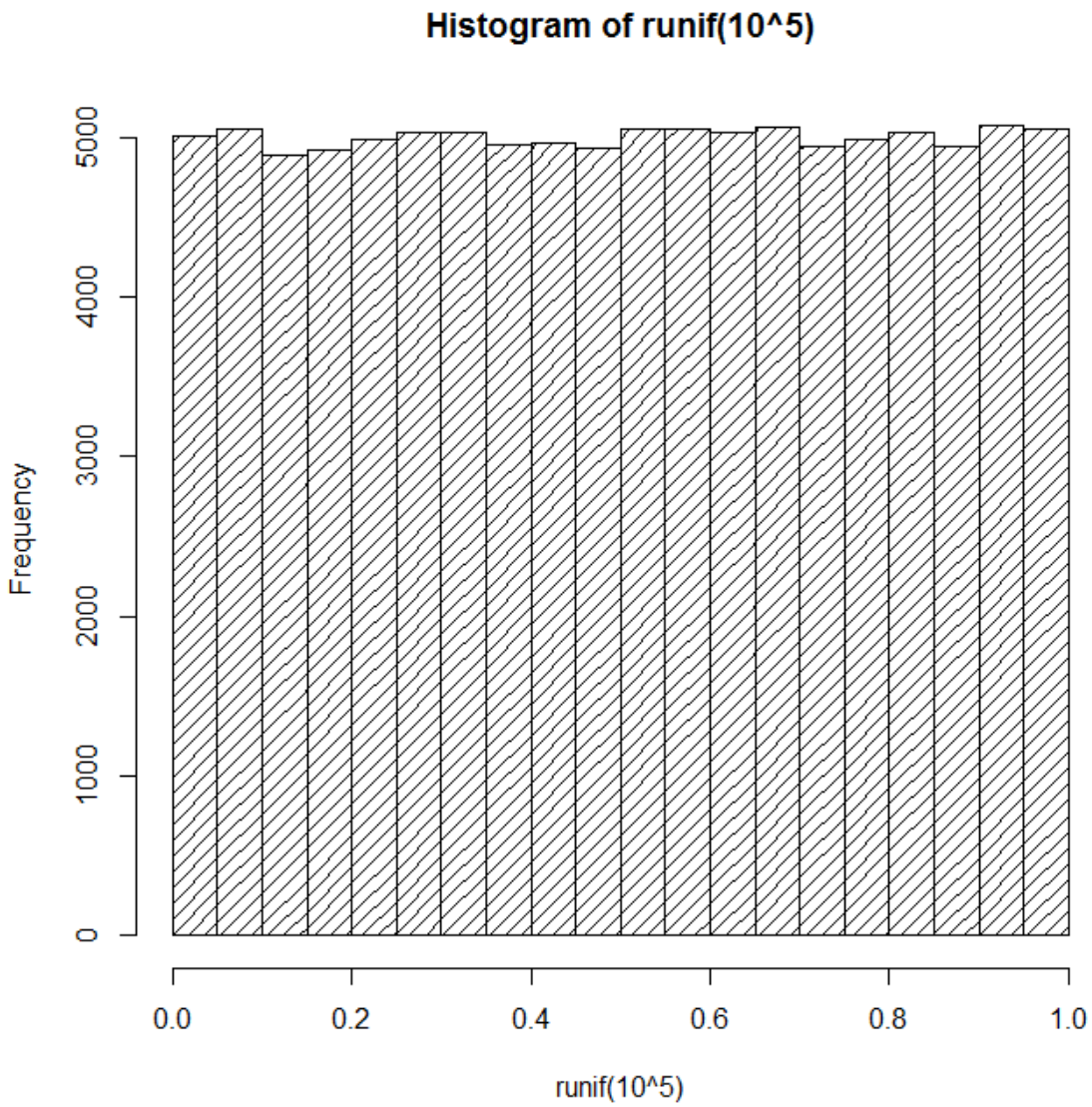
多重検定

- ほぼすべての帰無仮説が真の場合

たくさんの検定をすると、小さなp値がたくさん得られる

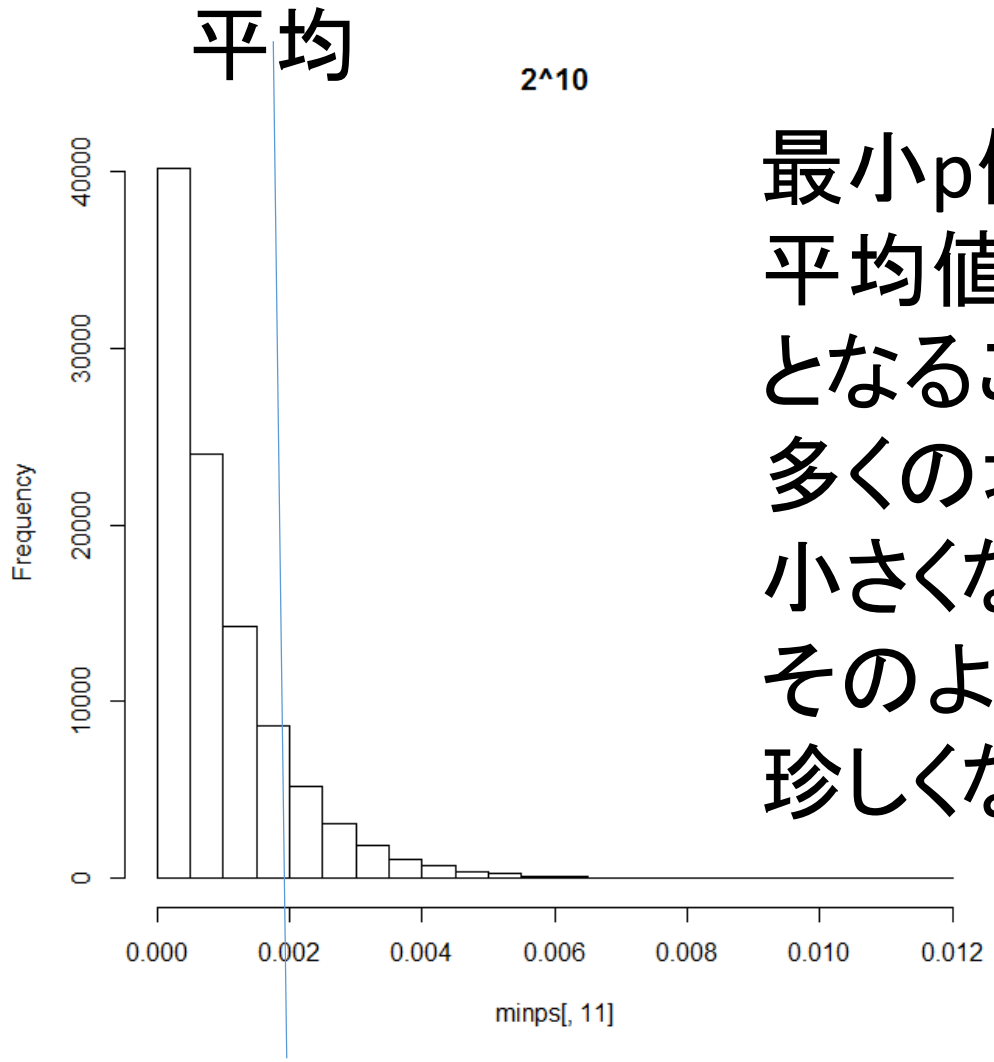
- １個の検定: 一様分布(0-1)
- 10個の検定 : 最小p値は0に近くなる、0.1くらい
- 100個の検定 : 最小p値はもっと0に近くなる、0.01くらい
- …

一様分布


Histogram of runif(10^5)
Frequency
0
1000
2000
3000
4000
5000
0.0
0.2
0.4
0.6
0.8
1.0
runif(10^5)

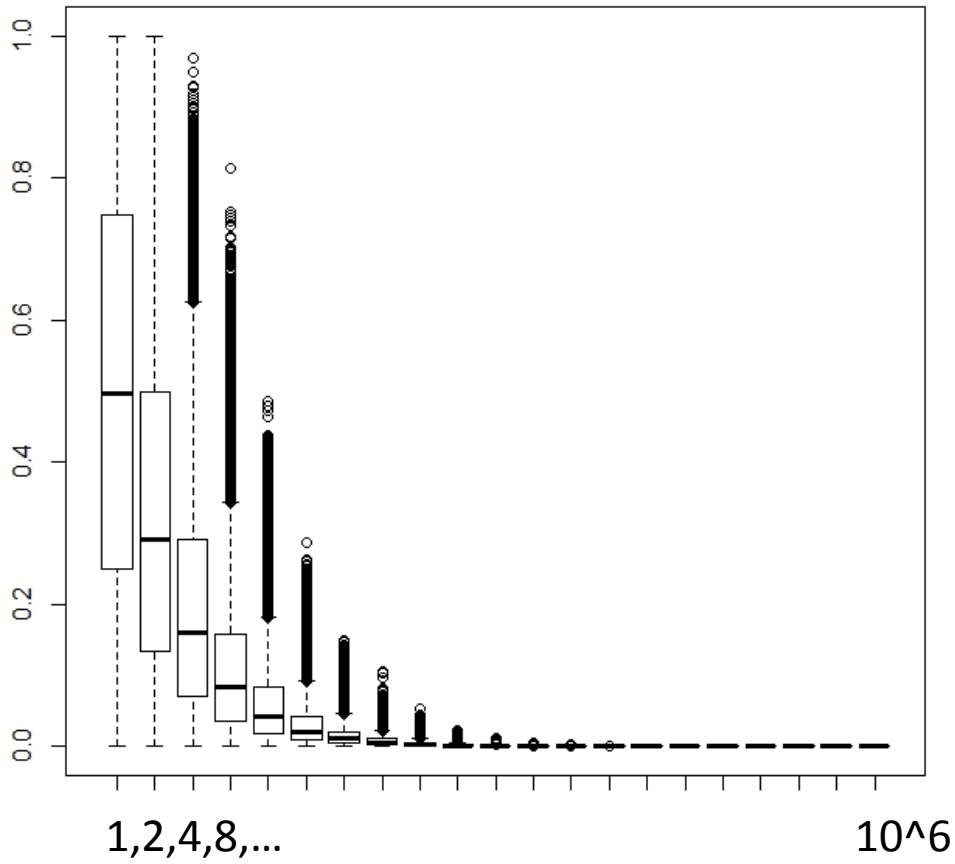
最小p値はどのように分布するか

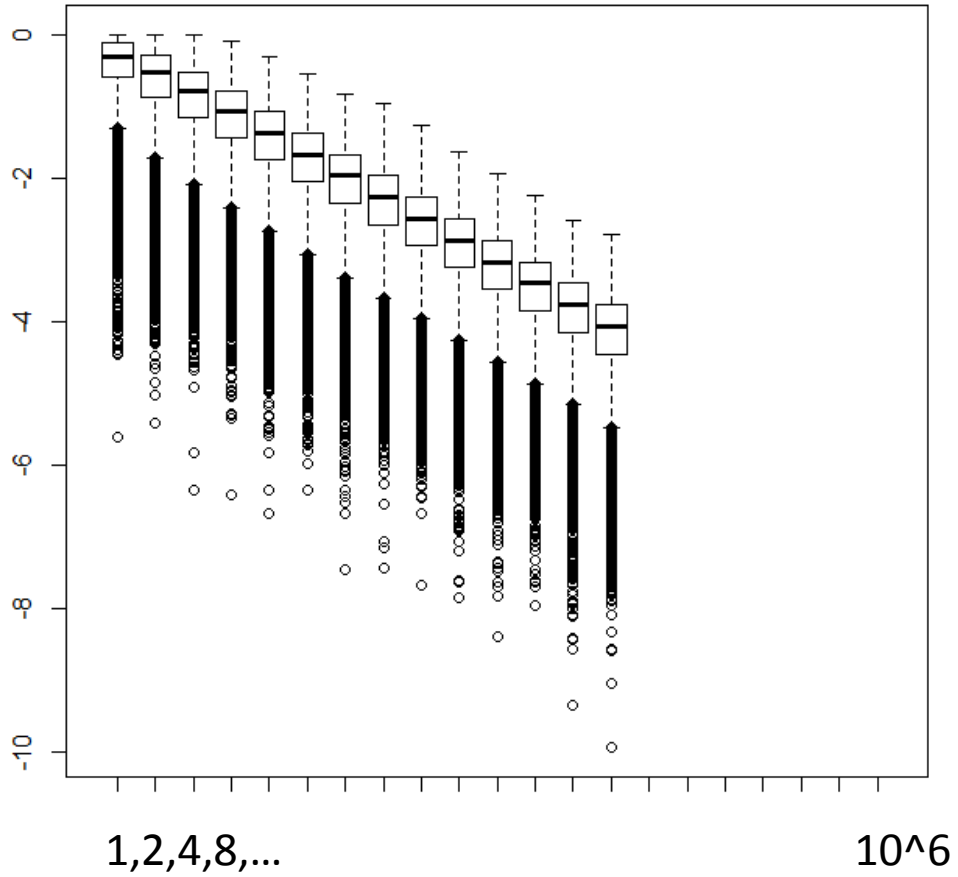
- 2^10



最小p値が
平均値よりかなり大きな値
となることもあるが、
多くの場合は、平均値より
小さくなる。
そのような小さなp値は
珍しくない。

最小p値の分布

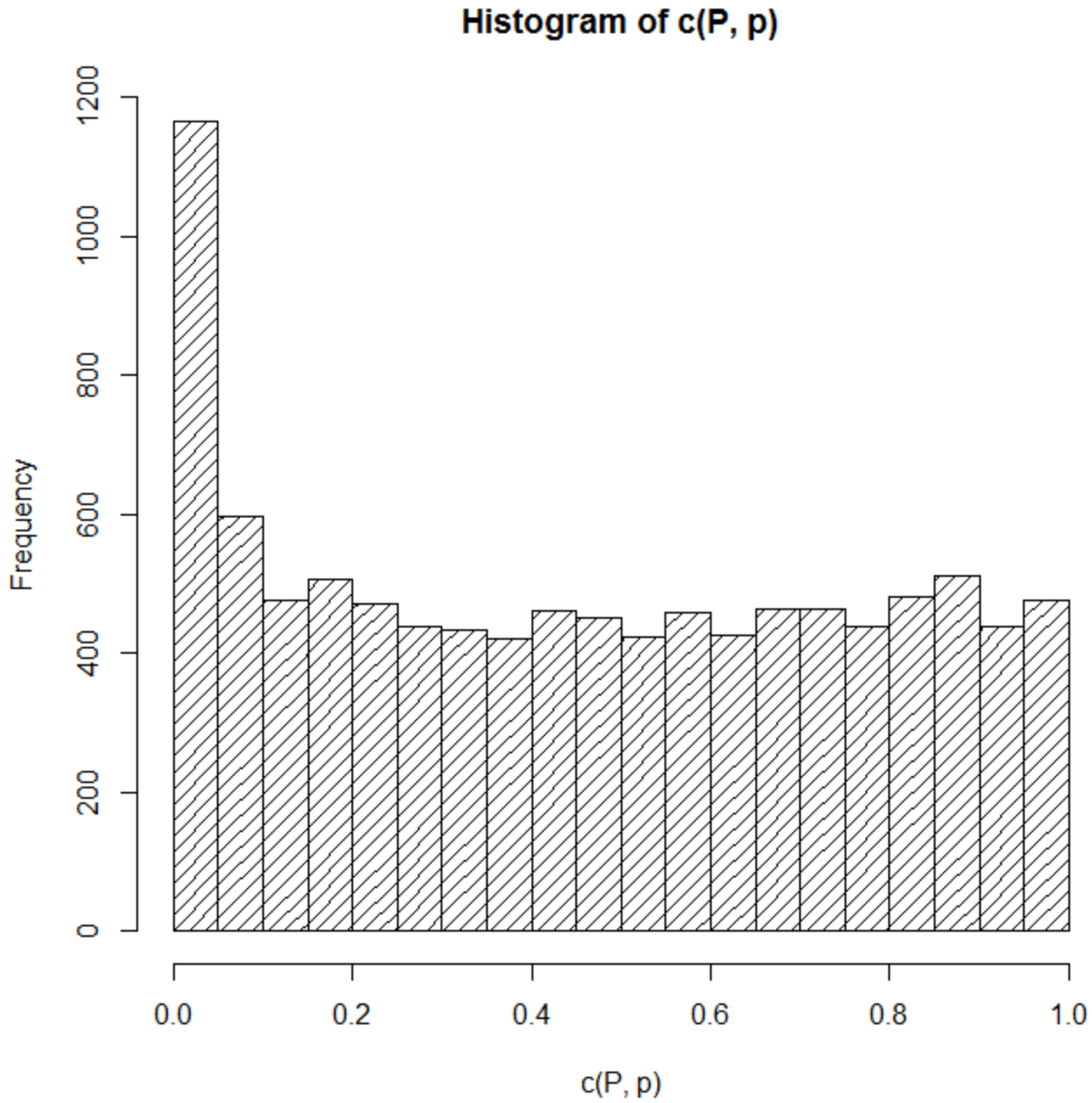




帰無が真でないとき, FDR (False Discovery Rate)

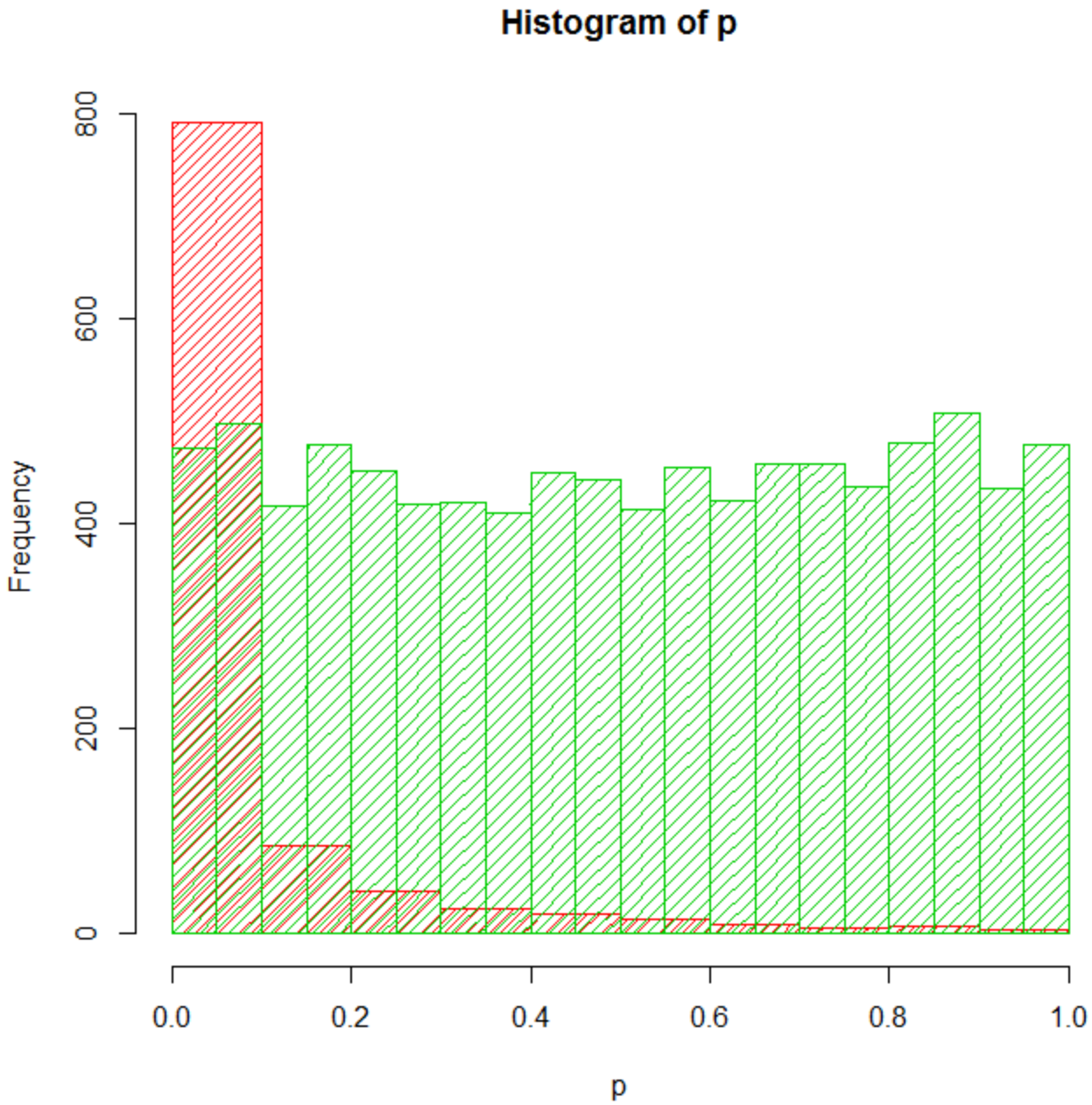
- 多数の仮説で帰無仮説が真でないとき、ほぼすべての仮説で帰無仮説が真でないとき

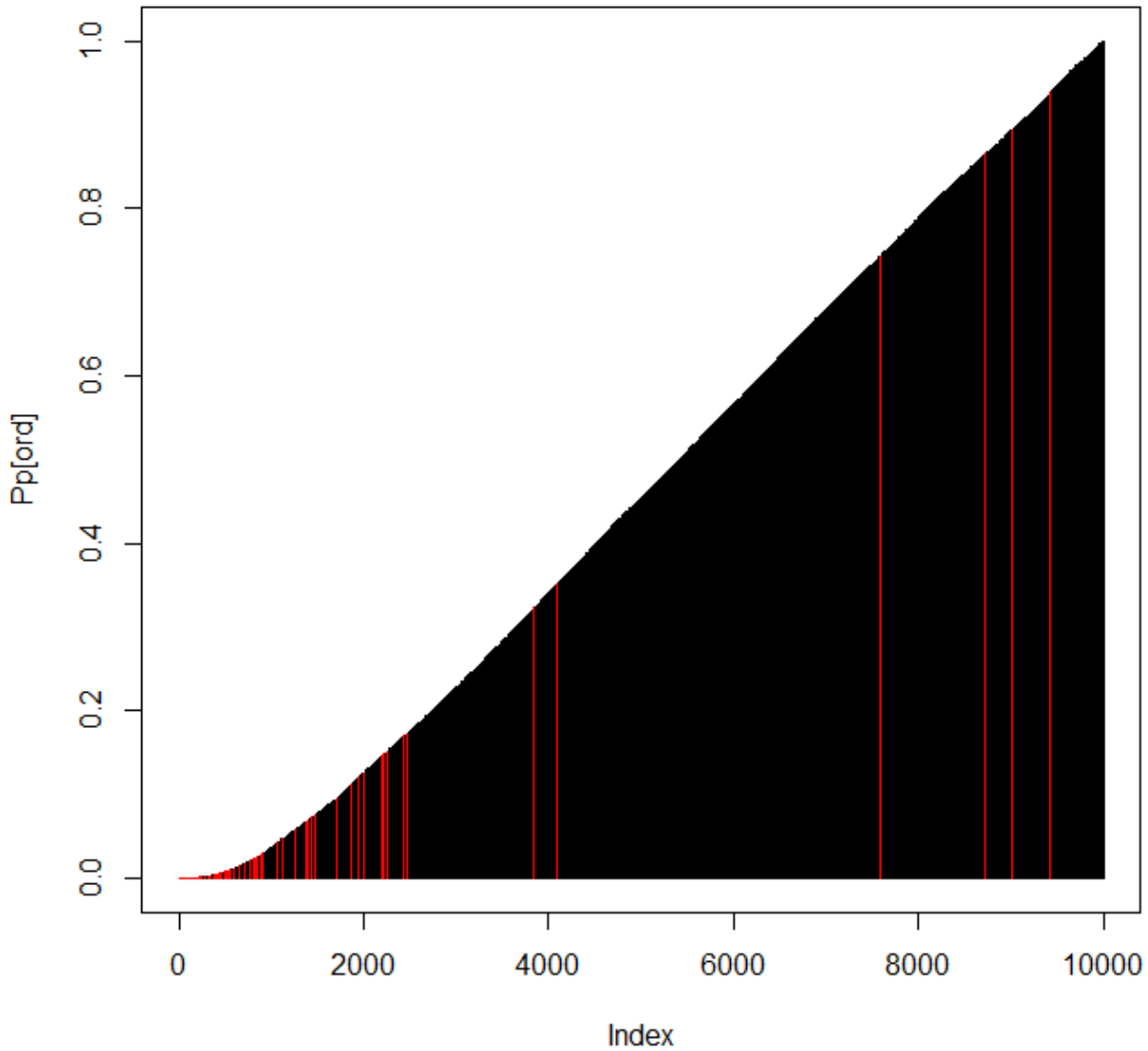
P-value

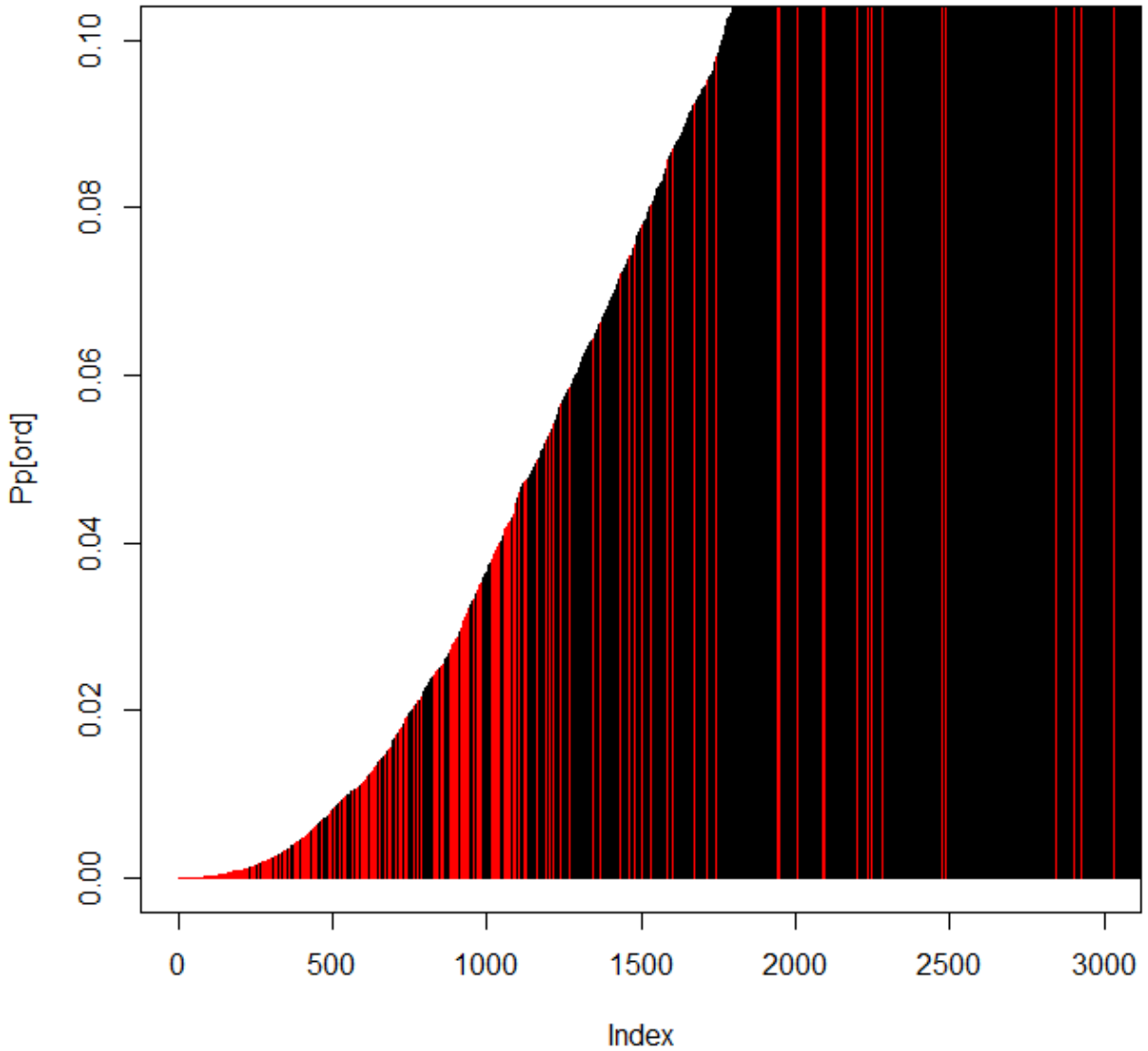

Histogram of c(P, p)
Frequency
0
200
400
600
800
1000
1200
0.0
0.2
0.4
0.6
0.8
1.0
c(P, p)

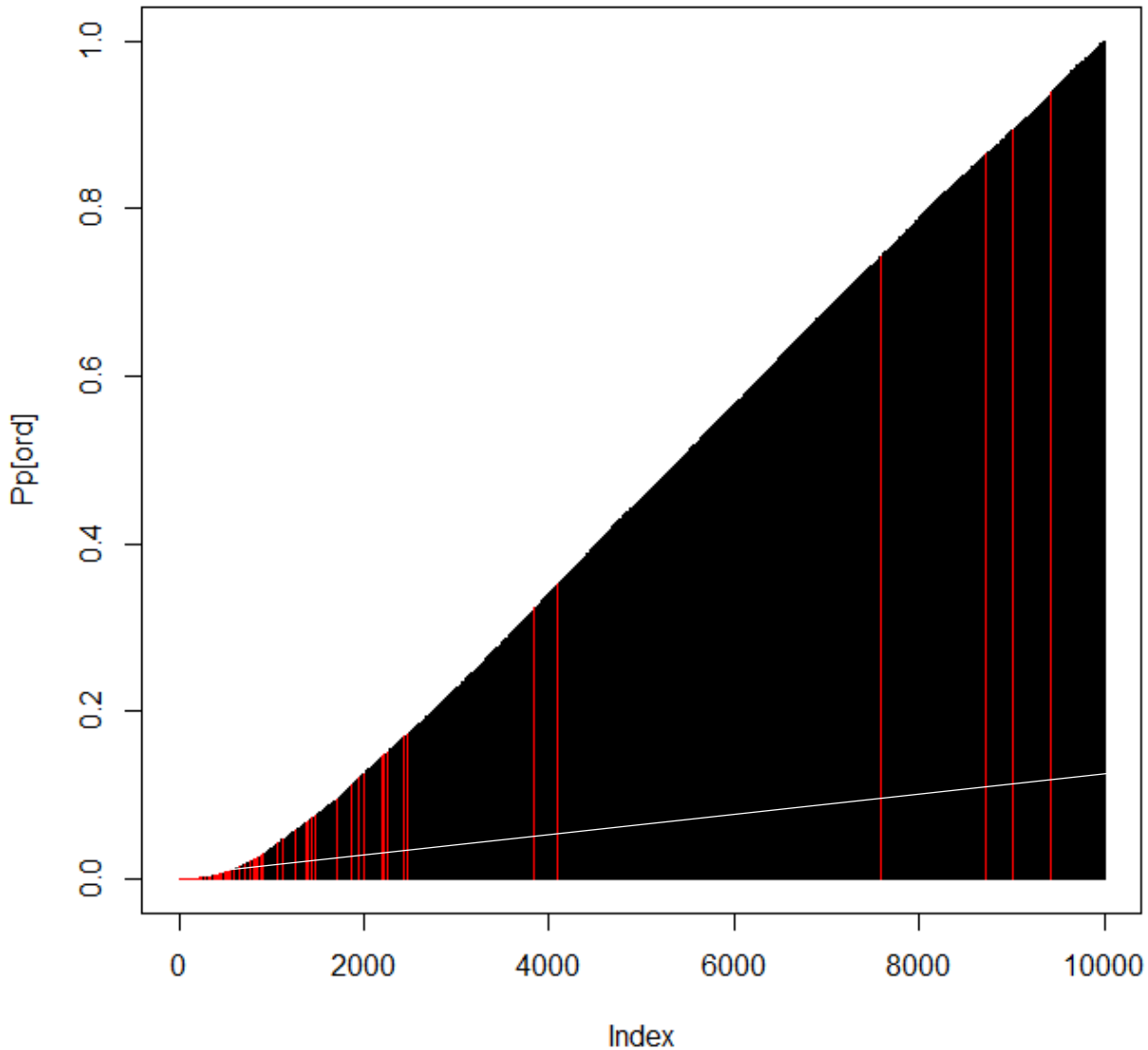
二つの分布を併せた分布

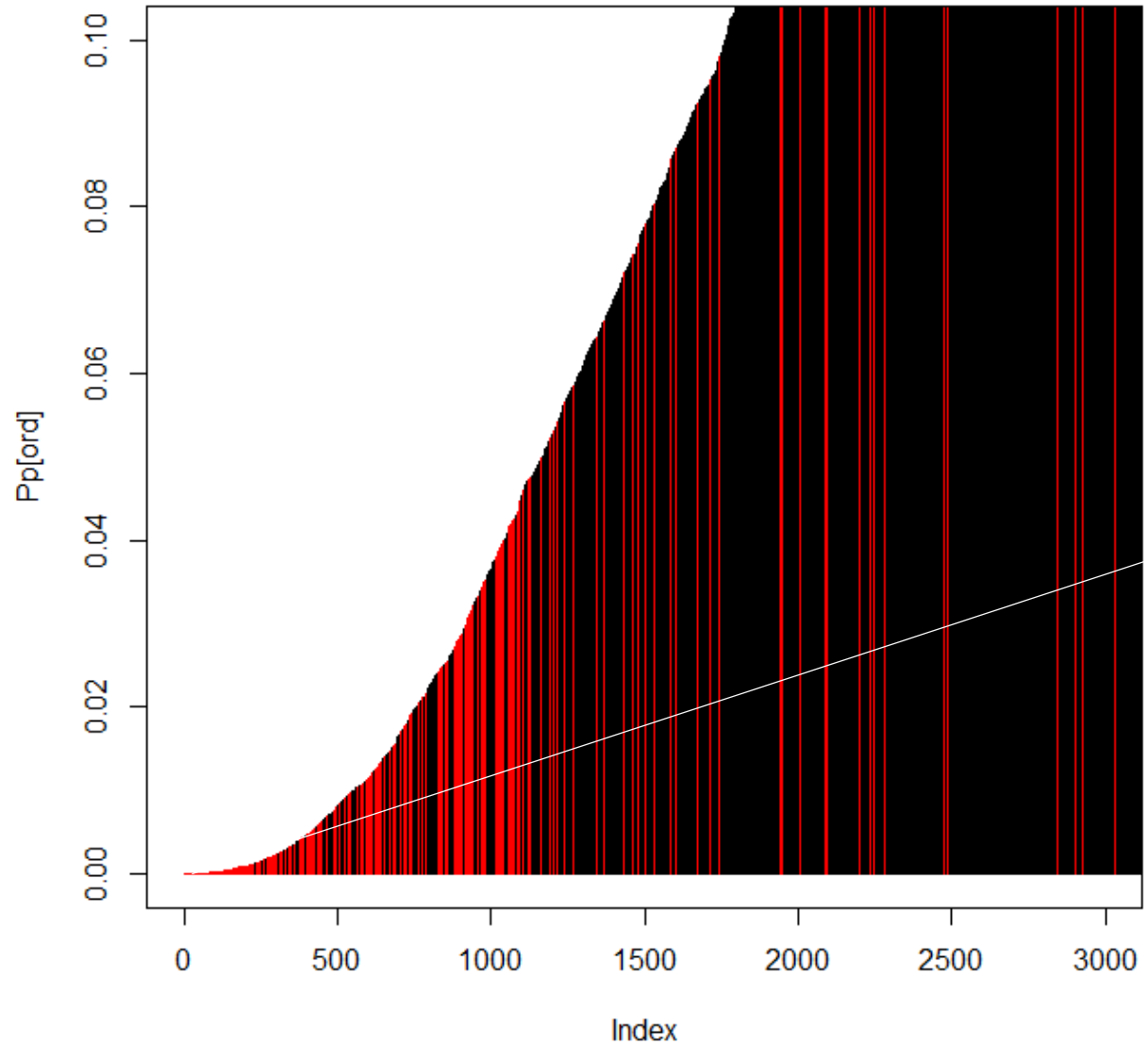
- 一様p値分布
- 小さ目のp値の分布

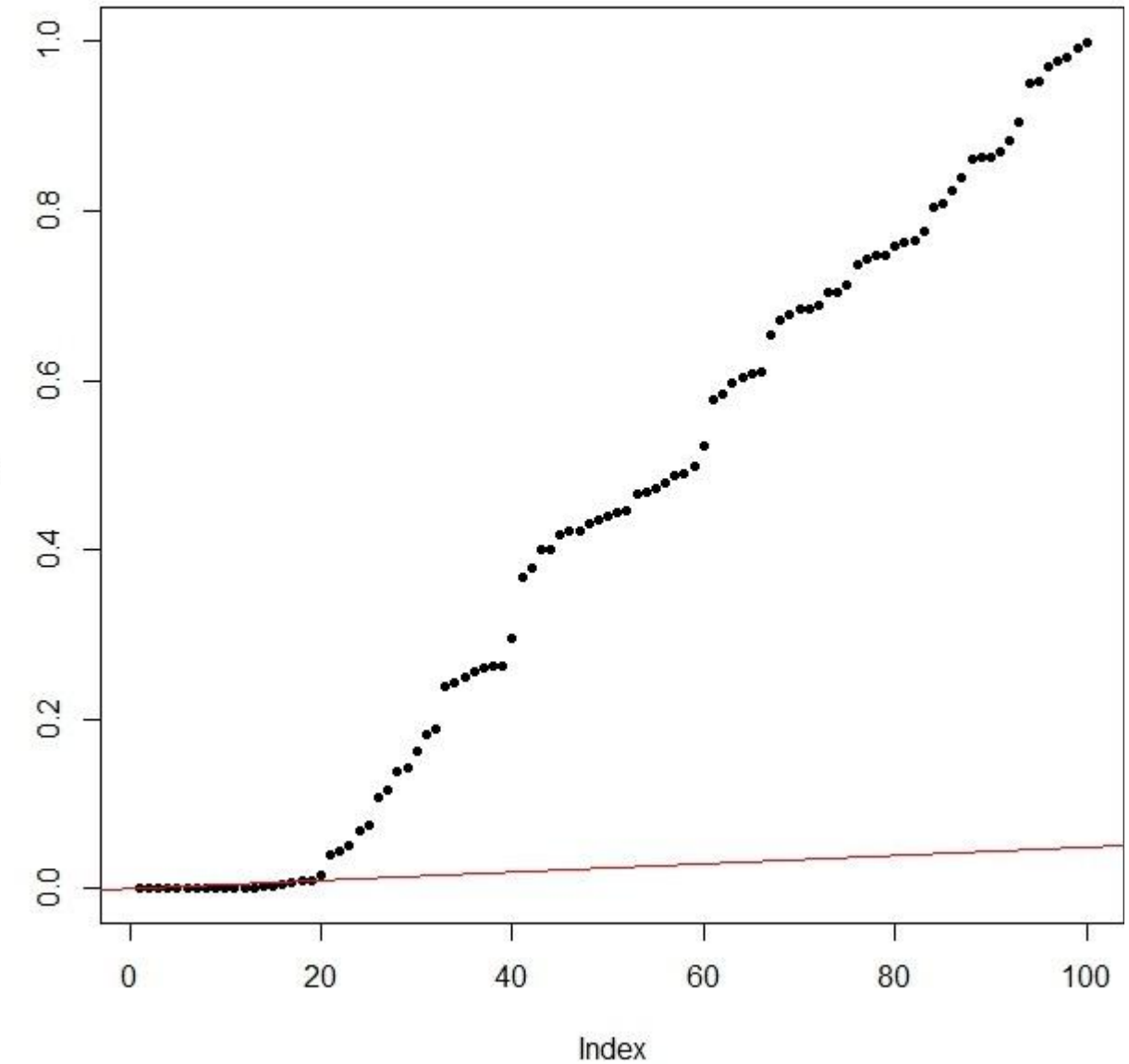


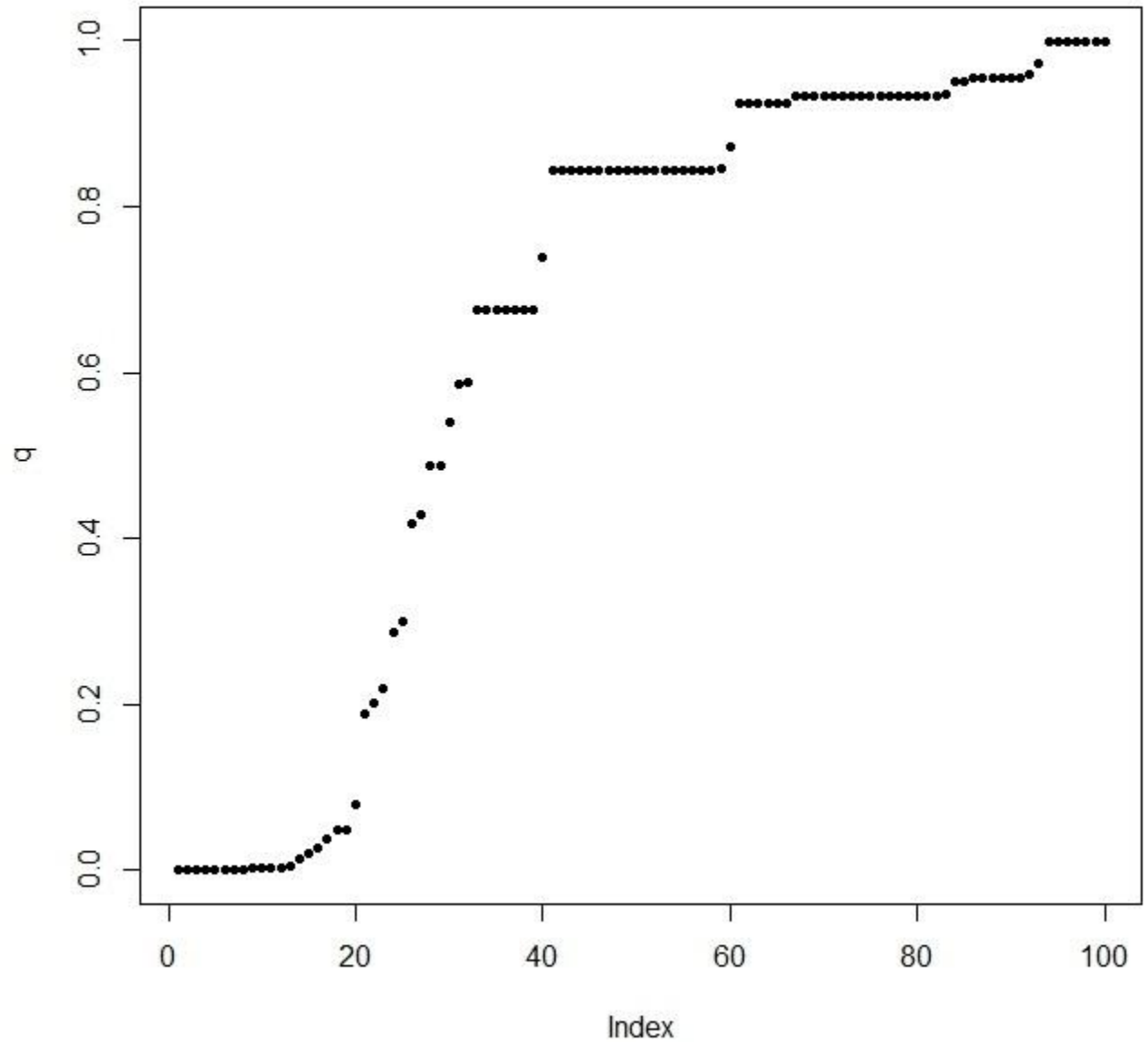

Pp[ord]
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
Index


Pp[ord]
0.00
0.02
0.04
0.06
0.08
0.10
0
500
1000
1500
2000
2500
3000
Index

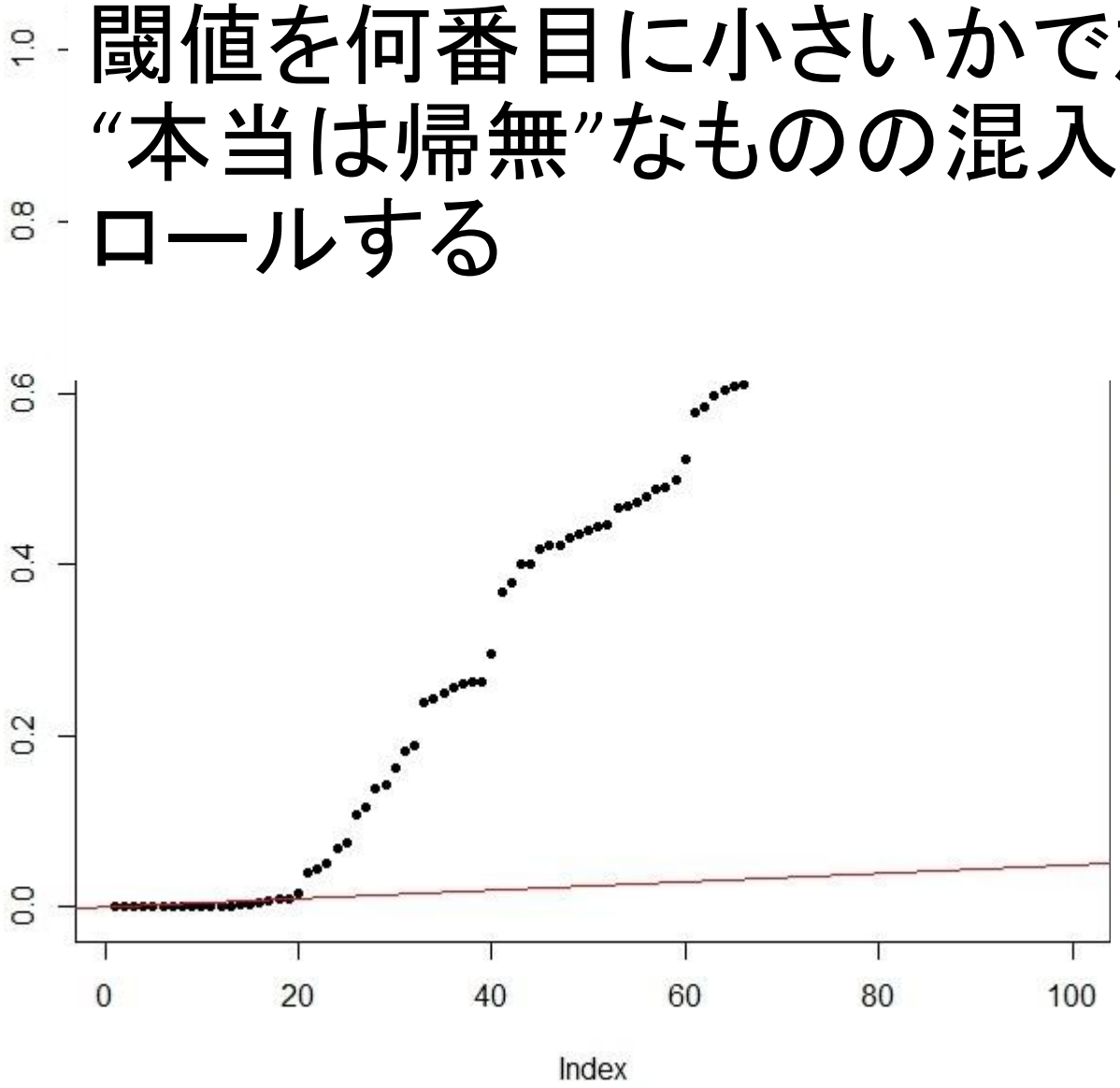

Pp[ord]
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
Index

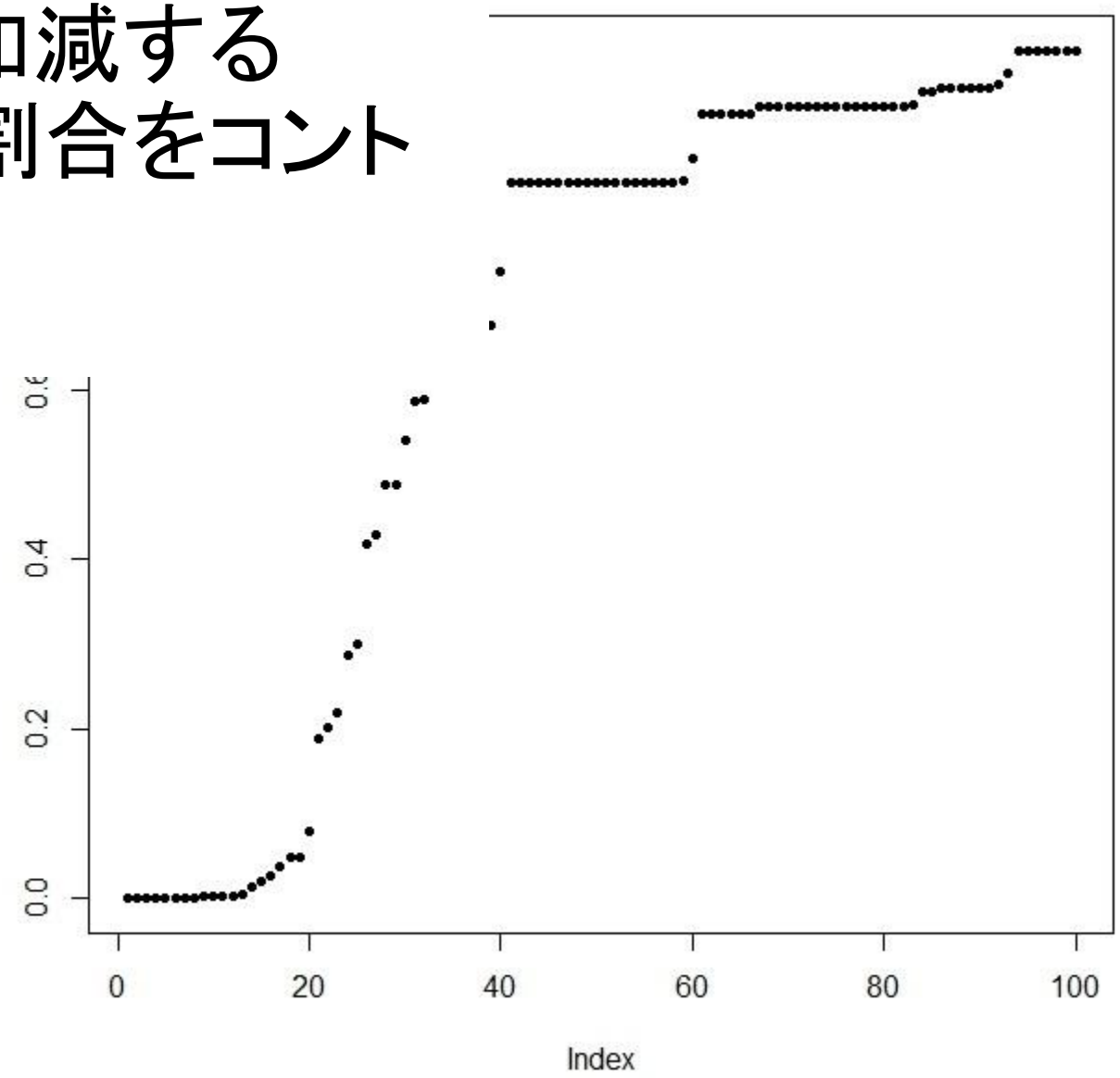

Pp[ord]
0.00
0.02
0.04
0.06
0.08
0.10
0
500
1000
1500
2000
2500
3000
Index


0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Index


q
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Index

小さいものを拾う
閾値を何番目に小さいかで加減する
“本当は帰無”なものの混入割合をコントロールする

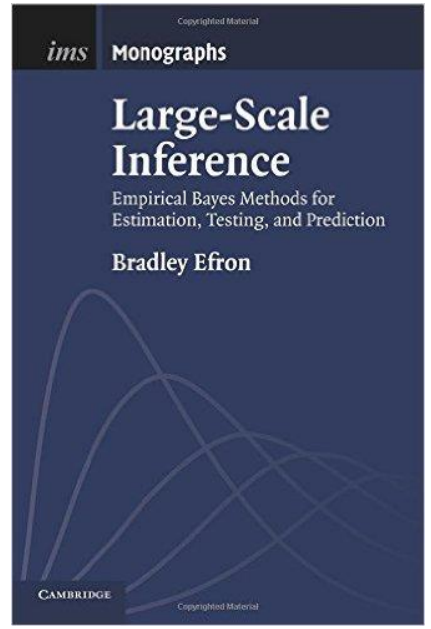




Large-scale inference

- たくさんのものを一度に測定したら、その分布には意味がある
- 分布を活用すると、個々の対象の推定値は、単独での推定値と変わってくる
- FDRのQ値もそんな枠組み

- 「一度に観測した多数が作る分布」を使う～経験ベイズ～



検定・推定・分類

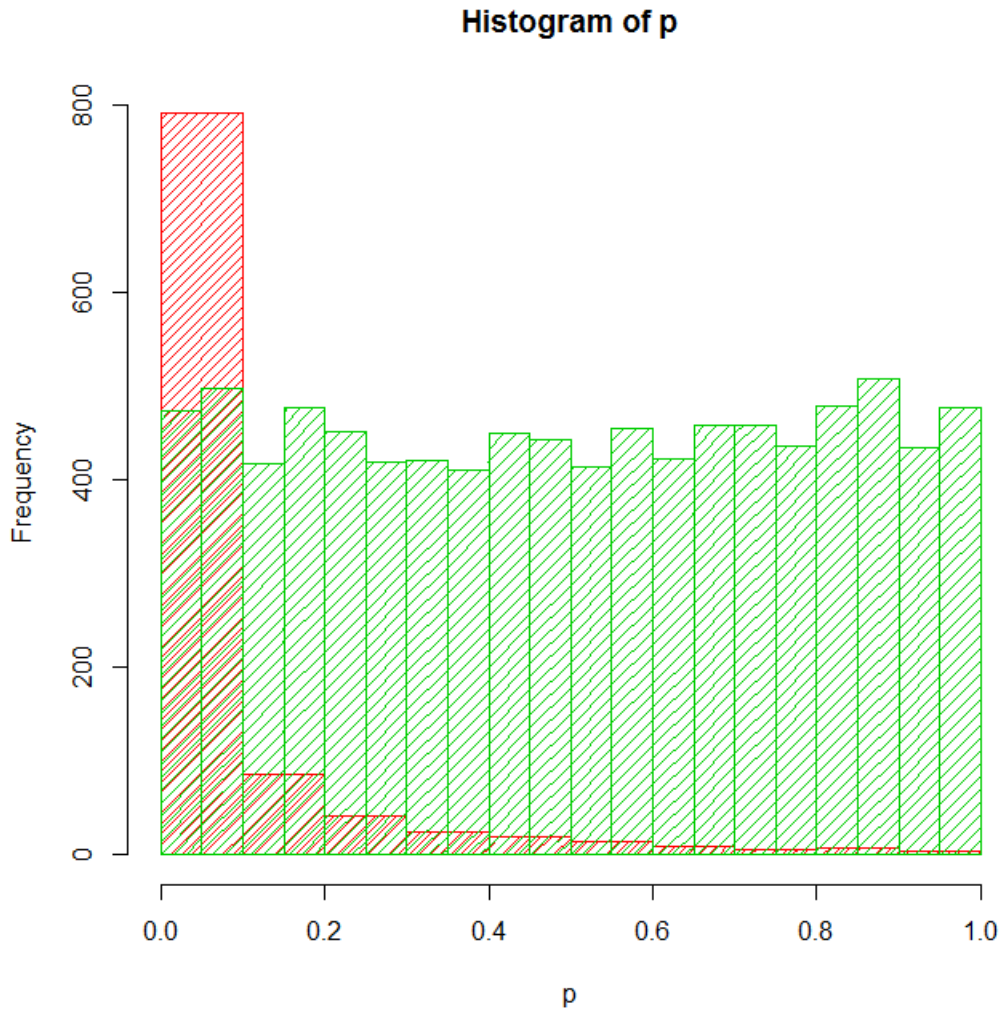
- 検定
  - 有意、エラーのコントロール、多重検定
- 推定
  - 区間推定、モデル推定、ベイズ
- 分類
  - 教師アリ、教師ナシ

推定

- モデル、パラメタ、区間推定、ベイズ

- 一様p値分布
- 小さ目のp値分布

この2色分け、という想定はモデル



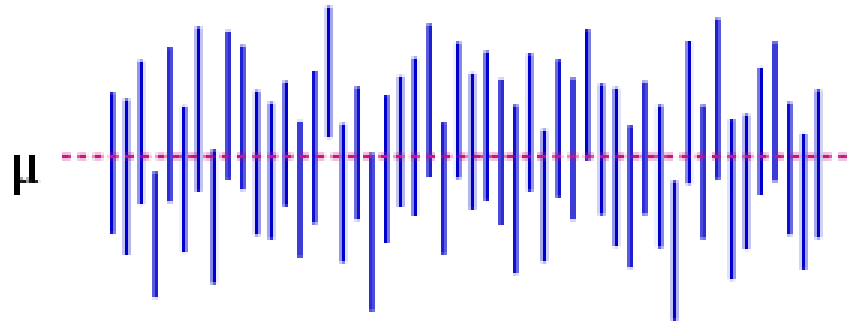
推定

- サンプル　→　点推定、信頼区間(区間推定)

- 標本分布、理論的な推定値、不偏推定値...

推定

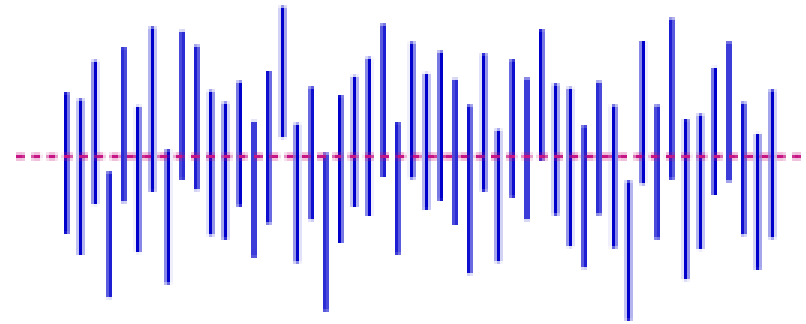
- サンプル　→　点推定、信頼区間(区間推定)
- 標本分布、理論的な推定値、不偏推定値…



「『海王星の質量は a から b の間である』といえば、10回に9回くらいは当たっているだろう」

推定

- サンプル　→　点推定、信頼区間(区間推定)
- 標本分布、理論的な推定値、不偏推定値…
- 頻度主義



「『海王星の質量は a から b の間である』といえば、10回に9回くらいは当たっているだろう」

推定

- 頻度主義　vs.　ベイジアン

- 頻度主義(である有意性検定)の理論体系は、その利用者に不自然な思考を強いるからです。また数学的に高度であり、文科系の学生には理解ではなく、暗記を強いるからです。
- 対して研究仮説が正しい確率を直接計算するベイズ流の推論は考え方がとても自然です。

推定

- ベイジアン
- モデルにはパラメタ
- データ　＋　モデル　→　パラメタの値の推定
- 推定には尤度。最尤推定。尤度に基づく区間推定


データ ＋ モデル
↓
パラメタの値の推定

まとめ：ジェノタイプ・フェノタイプという値

データ ＋ モデル
↓
パラメタの値の推定

- データ解析するために
  - 「値」として取り出す
  - 「値」にも色々
  - いわゆる「値」とは、「数」
  - 「数」とは
    - 自然数・整数・有理数・実数・複素数・ベクトル・行列…
  - いわゆる「値」ではない、データ解析用の「値」とは
    - 数理モデル
      - 特に、自然現象では、ばらつきがあることが基本なので
      - 確率モデル・統計モデル
    - ただし、モデルは(広義の)パラメタで構成するので
    - パラメタの「値」を扱うと言う意味では、「数」に戻る

  - 「いわゆる値」は単純な数理・確率モデルでのパラメタ値
  - より複雑な「タイプ」は複雑なモデルでのパラメタ値

推定

- 頻度主義　vs.　ベイジアン

- どちらか片方ではなく、両方使うのが、「今風」

- ベイジアンが目立つ理由
  - 込み入っているから・・・必然的事情
  - 計算機が使えるようになったから・・・複雑な分布でもシミュレーションで対処
  - データが大規模になったから・・・経験ベイズ

推定

- 頻度主義　vs.　ベイジアン

- ノイズのあるハイスループットデータのテ
- 検定・推定・分類
- 多次元・高次元データ
- 乱数を使ったアプローチ
- その他：実験デザイン

- どちらか片方ではなく、両方使うのが、「今風」

- ベイジアンが目立つ理由
  - 込み入っているから・・・必然的事情
  - 計算機が使えるようになったから・・・複雑な分布でもシミュレーションで対処
  - データが大規模になったから・・・経験ベイズ

Estimation/Inference

- 頻度主義　vs.　ベイジアン

- ノイズのあるハイスループットデータのテ
- 検定・推定・分類
- 多次元・高次元データ
- 乱数を使ったアプローチ
- その他：実験デザイン

- どちらか片方ではなく、両方使うのが、「今風」

- ベイジアンが目立つ理由
  - 込み入っているから・・・必然的事情
  - 計算機が使えるようになったから・・・複雑な分布でもシミュレーションで対処
  - データが大規模になったから・・・経験ベイズ

Estimation/Inference

- ノイズのあるハイスループットデータのテ
- 検定・推定・分類
- 多次元・高次元データ
- 乱数を使ったアプローチ
- その他：実験デザイン

- 頻度主義　vs.　ベイジアン

- どちらか片方ではなく、両方使うのが、「今風」

- ベイジアンが目立つ理由
  - 込み入っているから・・・必然的事情
  - 計算機が使えるようになったから・・・複雑な分布でもシミュレーションで対処
  - データが大規模になったから・・・経験ベイズ

Estimation/Inference

- 頻度主義　vs.　ベイジアン

- ノイズのあるハイスループットデータのテ
- 検定・推定・分類
- 多次元・高次元データ
- 乱数を使ったアプローチ
- その他：実験デザイン

- どちらか片方ではなく、両方使うのが、「今風」

- ベイジアンが目立つ理由
  - 込み入っているから・・・必然的事情
  - 計算機が使えるようになったから・・・複雑な分布でもシミュレーションで対処
  - データが大規模になったから・・・経験ベイズ

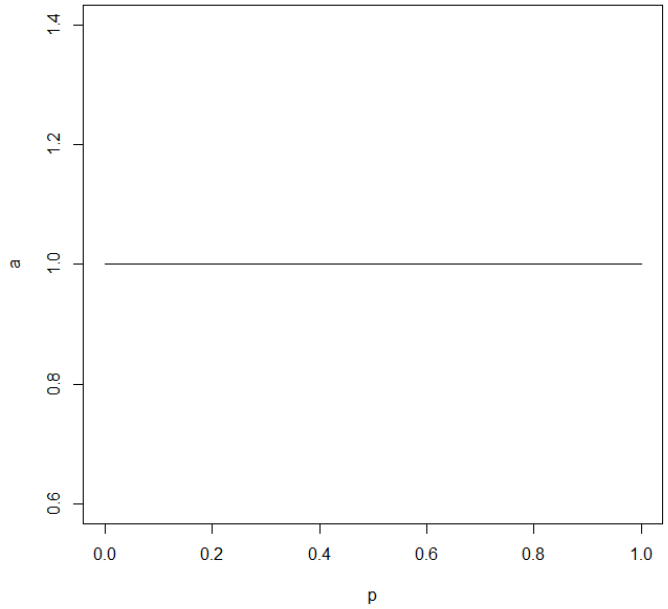
推定

- 頻度主義　vs.　ベイジアン

- 「事前分布」がないと使えない
- 「正しい事前分布」とは何か…

成功率：その、情報なしのときの事前確率

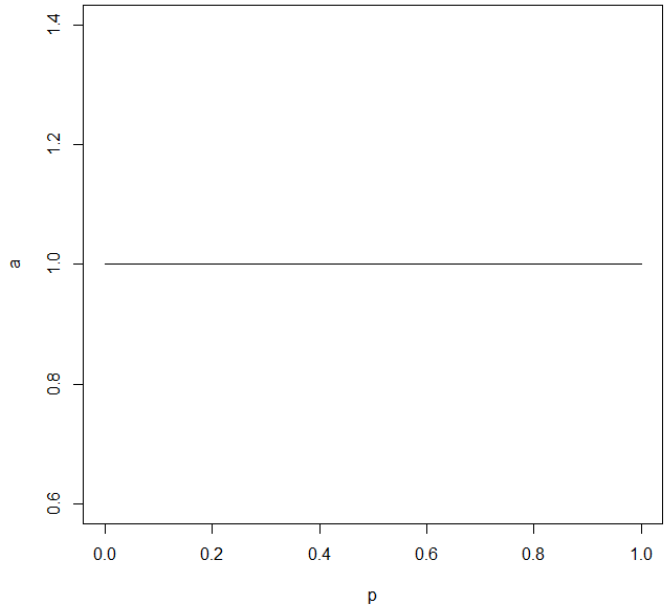
- 難易度も平均合格率も一切不明な、「変な資格試験」を、あなたが全く知らない「だれか」が受験すると言う。この人が合格する確率はいくつだと思うか？

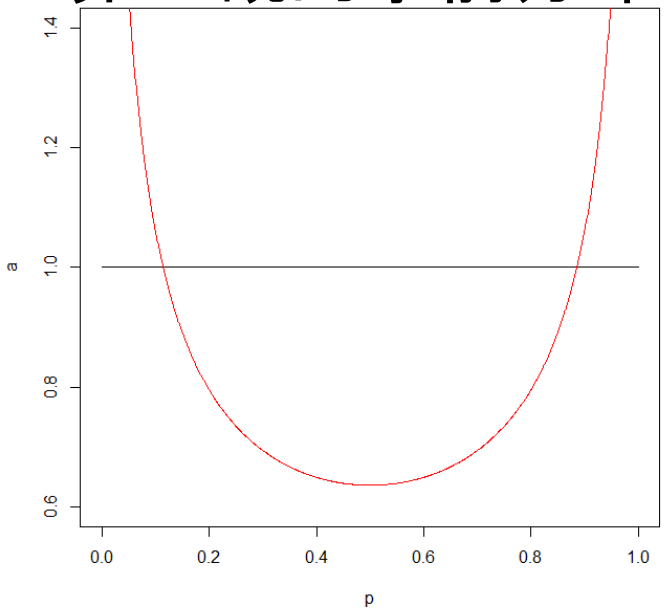


成功率：その、情報なしのときの事前確率

- 難易度も平均合格率も一切不明な、「変な資格試験」を、あなたが全く知らない「だれか」が受験すると言う。この人が合格する確率はいくつだと思うか？

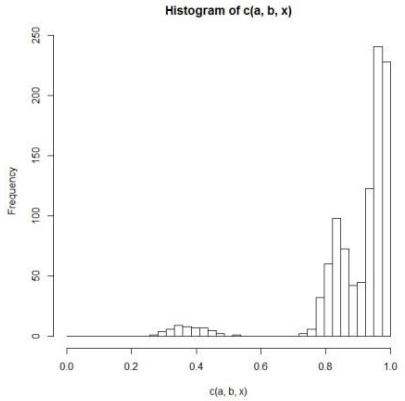
Jeffreys prior
非主観的事前分布の1つの取り方

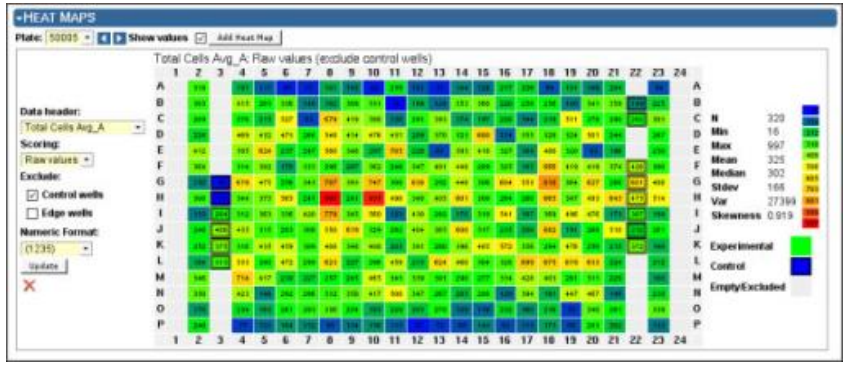




推定

- 頻度主義　vs.　ベイジアン
- どちらか片方ではなく、両方使うのが、「今風」
- 大規模データ Large scale inference : 経験ベイズは、取ったデータを活用した事前分布の設定





検定・推定・分類

- 検定
  - 有意、エラーのコントロール、多重検定
- 推定
  - 区間推定、モデル推定、ベイズ
- 分類
  - 教師アリ、教師ナシ

分類

- その前に、多次元/高次元 を

ゲノム・オミクス研究における統計・データサイエンスの役割

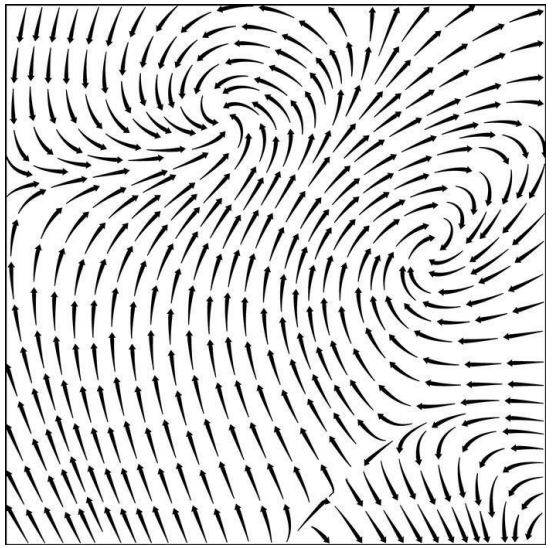
- ノイズのあるハイスループットデータのデータQC
- 検定・推定・分類
- 多次元・高次元データ
- 乱数を使ったアプローチ
- その他：実験デザイン

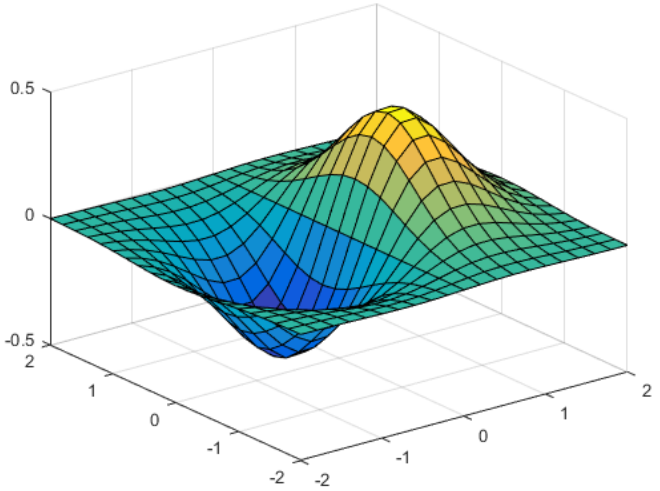
多次元・高次元データ

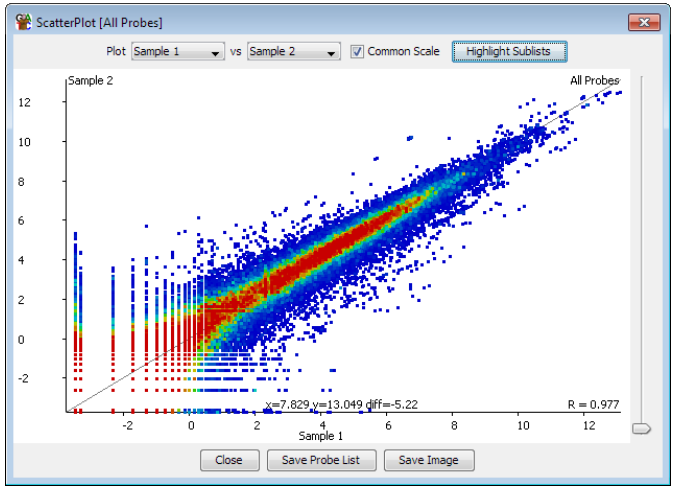
- 高次元データは「見られない」
- 高次元データをそのままの形で理解することはほぼ不可能

多次元・高次元データ

- 示せる次元はいくつまで？
- 空間は2次元か3次元
- それ以外の次元は
  - グレースケール、カラースケール
  - 矢印
  - 時間を使う(アニメーション)





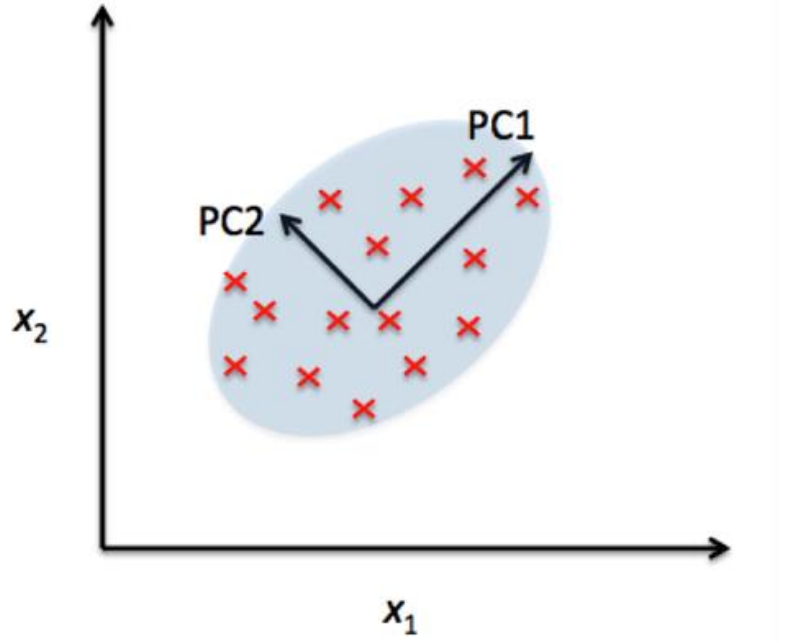


多次元・高次元データ

- 次元を下げる
  - 理解・視覚化可能な、重要な2，3の次元のみで切り取る

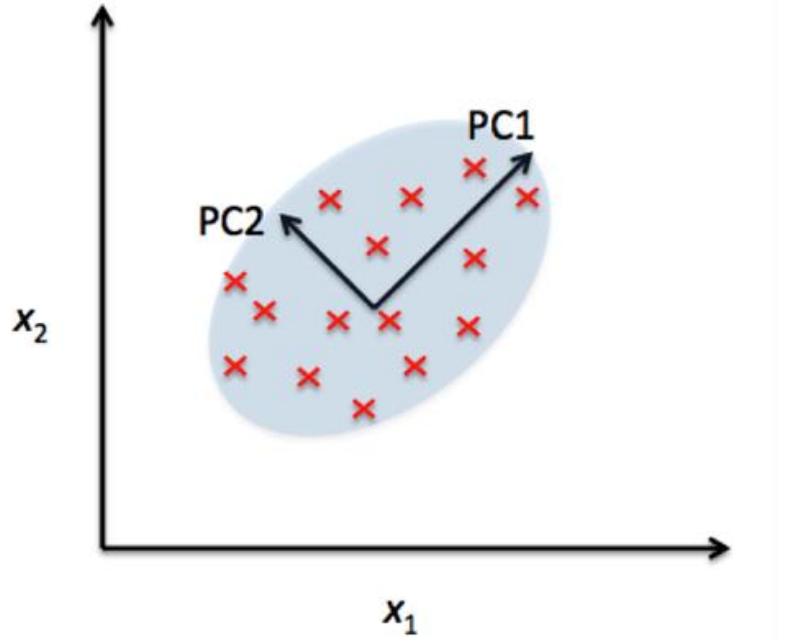
多次元・高次元データ

- 次元を下げる
  - 理解・視覚化可能な、重要な2，3の次元のみで切り取る
    - PCA (主成分分析)



多次元・高次元データ

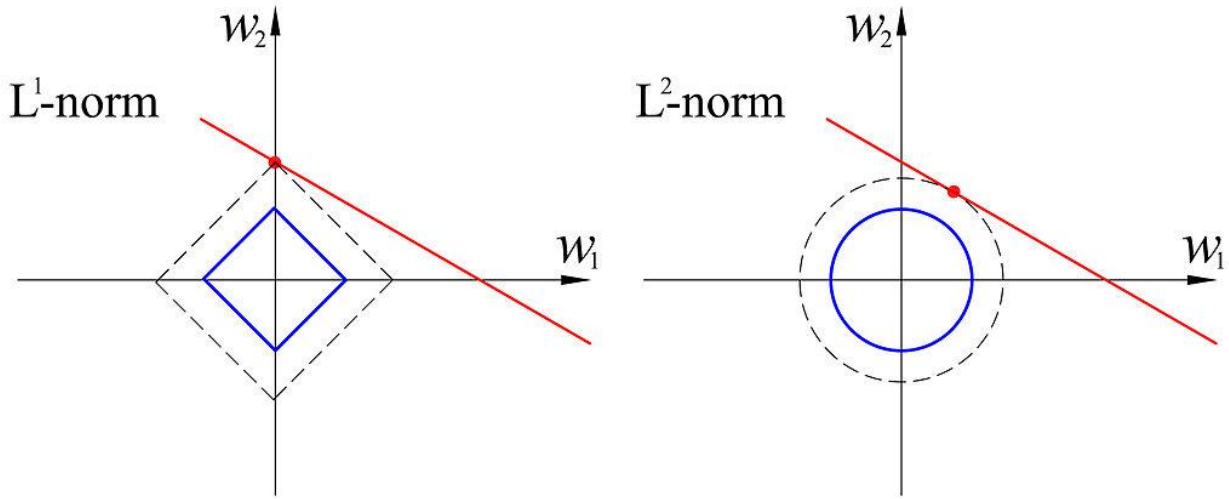
- 次元を下げる
  - 理解・視覚化可能な、重要な2，3の次元のみで切り取る
    - PCA (主成分分析)
  - そもそも、重要なのは、少次元なのであって、それ以外はノイズなので、切り取る

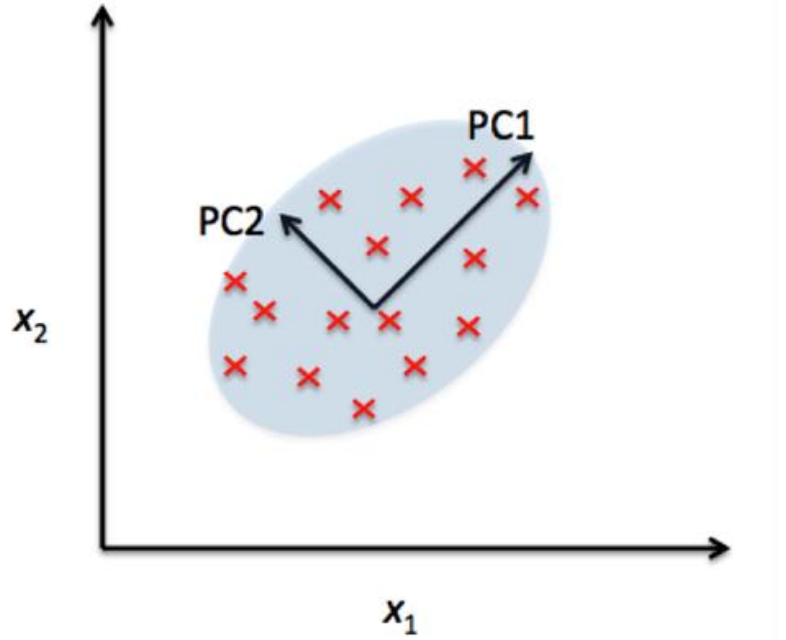


多次元・高次元データ

- 次元を下げる理解・視覚化可能な、重要な2，3の次元のみで切り取る
  - PCA (主成分分析)

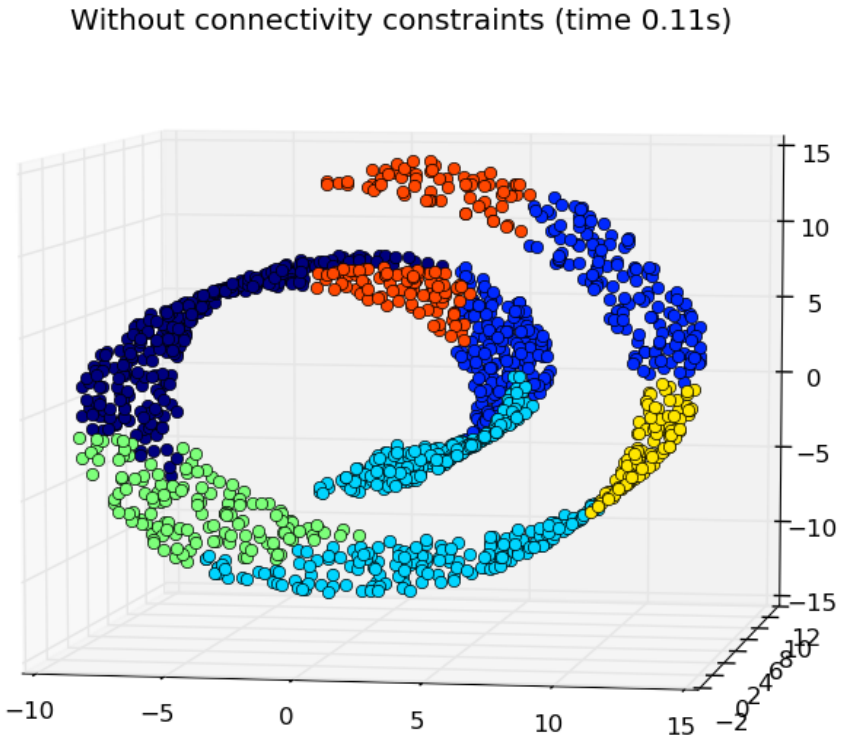
- そもそも、重要なのは、少次元なので、切り取る
  - LASSO, 圧縮センシング





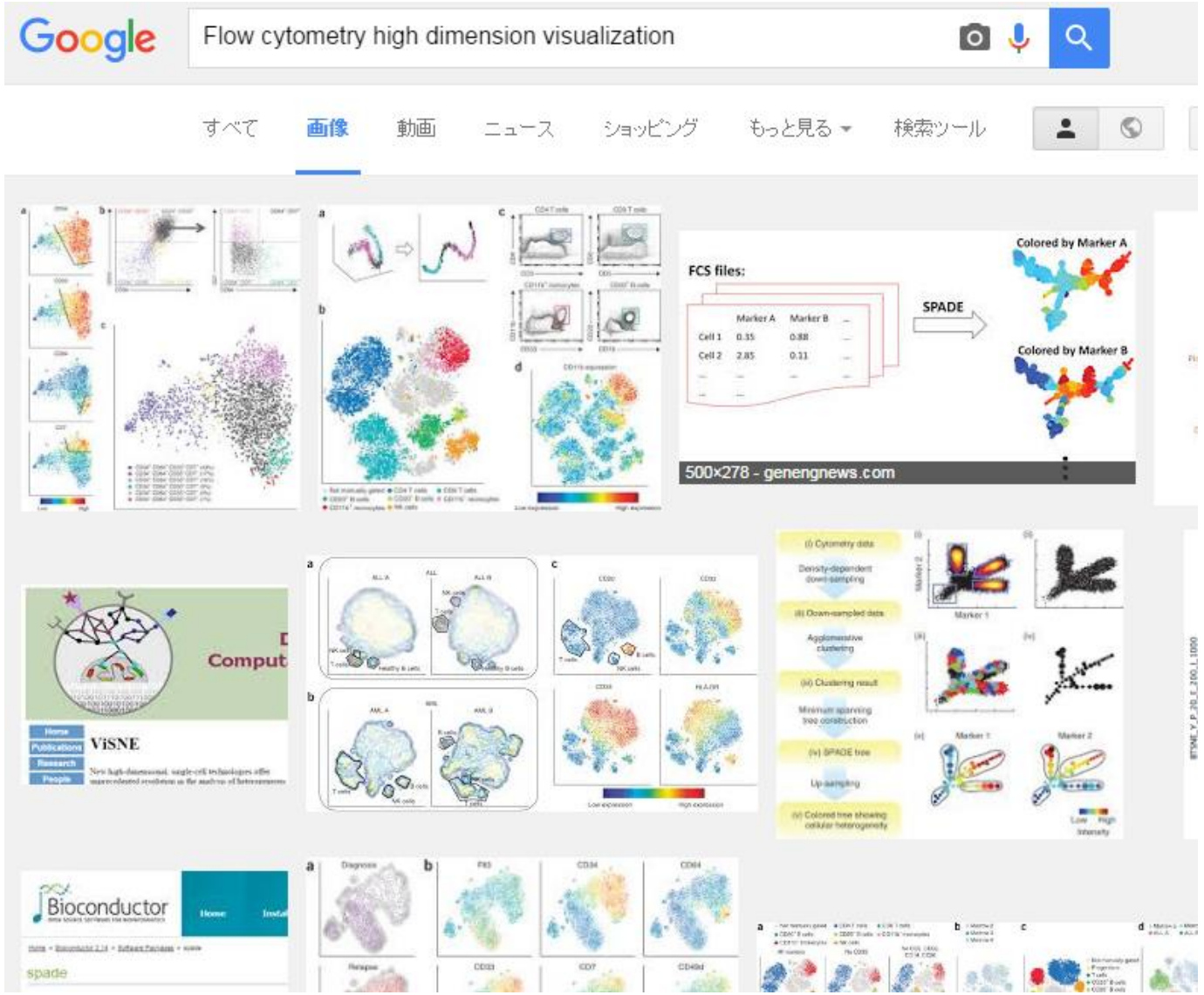
多次元・高次元データ

- 空間は高次元だが、データは低次元
- 多様体学習
- 高次元空間に投げ上げて、低次元に戻す



多次元・高次元データ

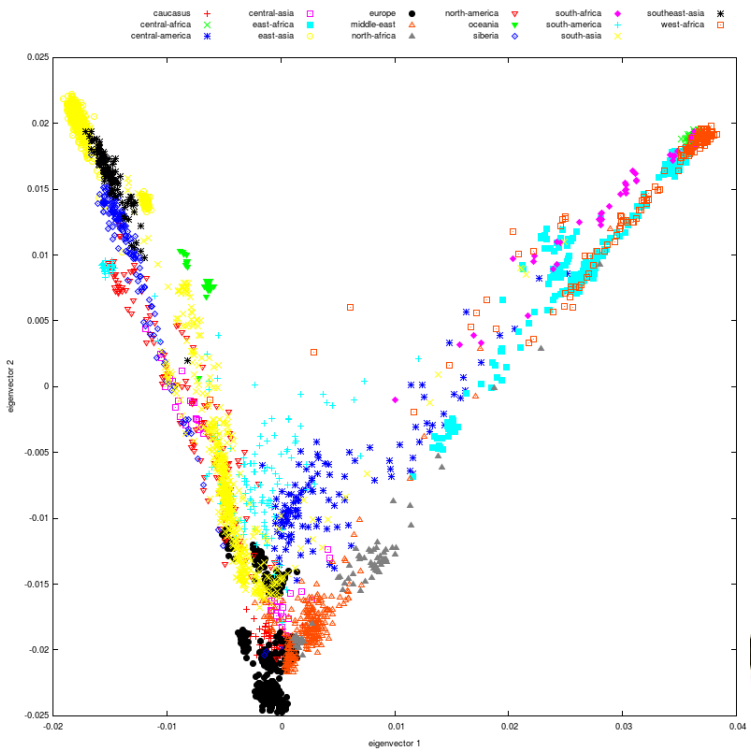
- たくさんの遺伝子
- たくさんのマーカー
- たくさんの観察項目



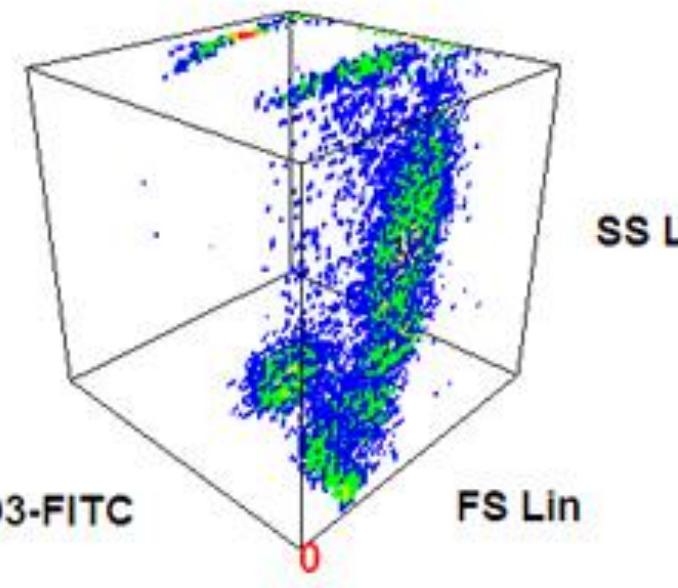
多次元・高次元データ

- ライフサイエンスデータは、高次元空間データとして観察されるが
  - 観察項目が膨大だから
- 項目間の類似・制約も大きく、思ったよりも低次元と、思われている

Ethnic diversity



FACS

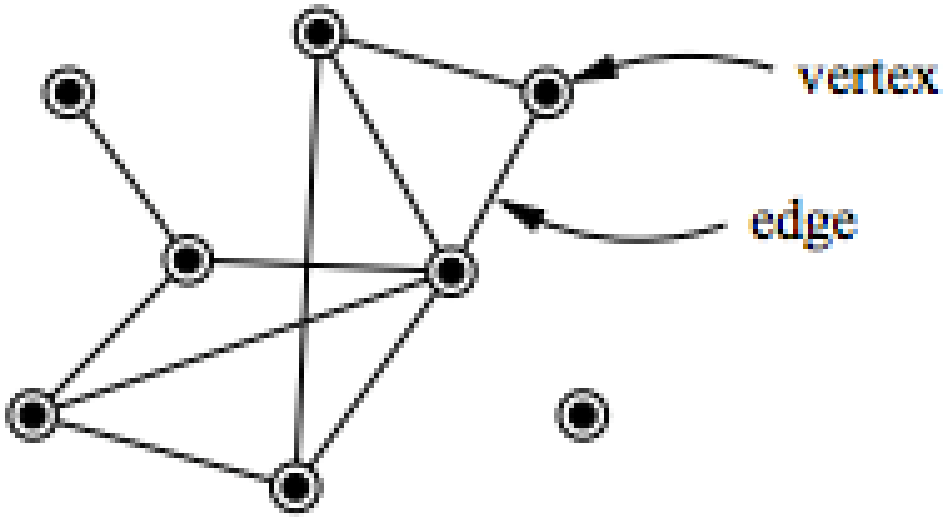


多次元・高次元データ

- 高次元空間の低次元オブジェクト～多様体～
- トポロジーを問題にする

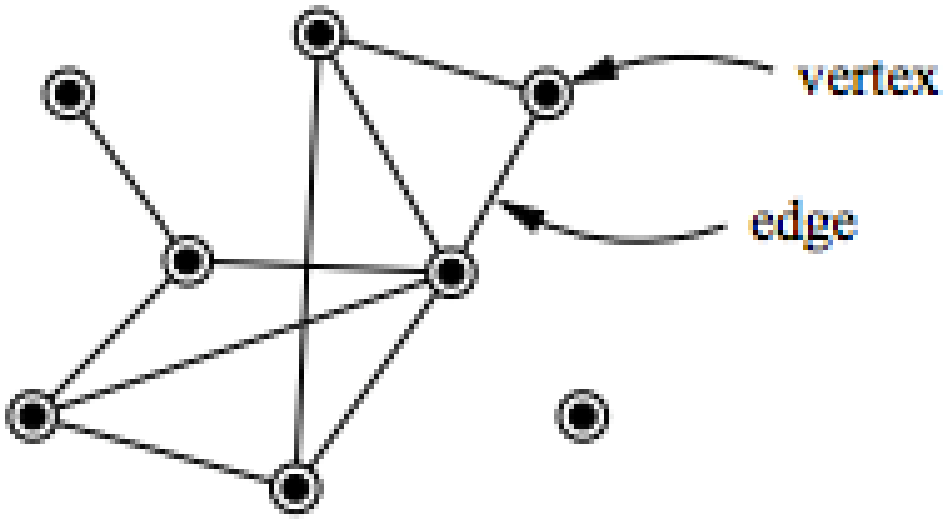
多次元・高次元データ

- 高次元間の低次元オブジェクト～多様体～
- トポロジーを問題にする
- グラフ・ネットワークとトポロジー



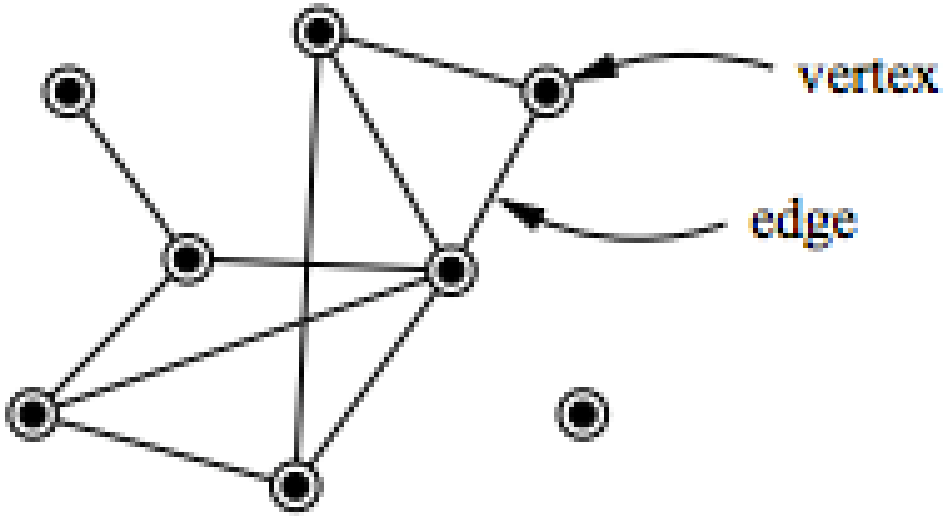
多次元・高次元データ

- グラフ：隣り合っていれば結ぶ
- 多要素のペア関係だけを考慮した単純化



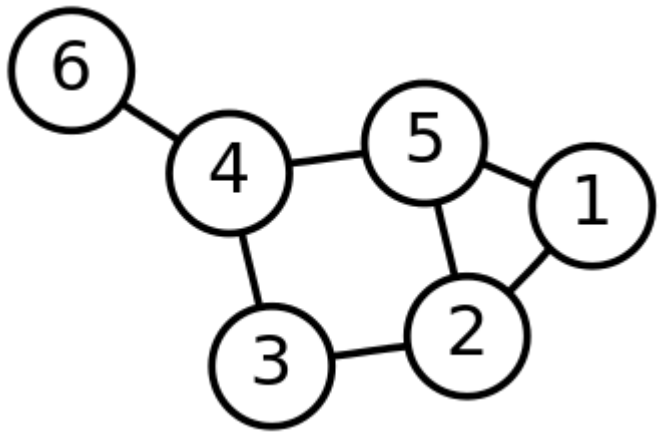
多次元・高次元データ

- グラフ：隣り合っていれば結ぶ
- 多要素のペア関係だけを考慮した単純化
- トリオ以上の組み合わせを無視した評価系



多次元・高次元データ

- グラフと線形解析

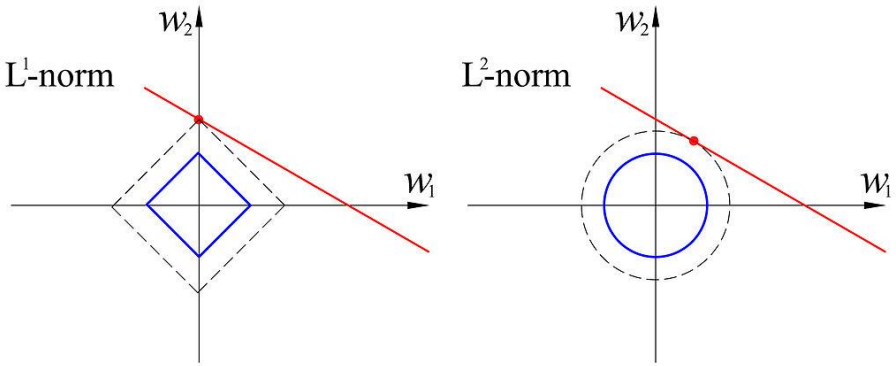


行列で表現できる。

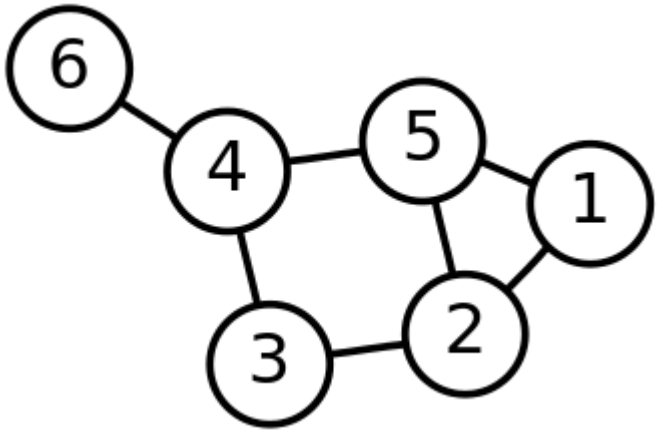
$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

多次元・高次元データ

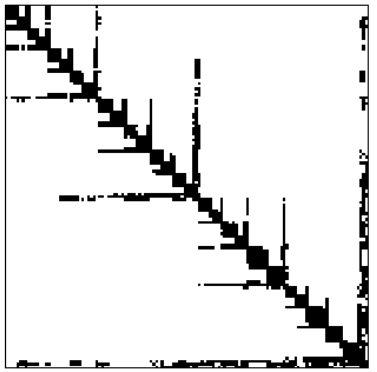
- グラフと線形解析



- グラフと疎解析



$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & & & \\ 0 & 0 & 1 & & & \\ 1 & 1 & 0 & & & \\ 0 & 0 & 0 & & & \end{pmatrix}$$



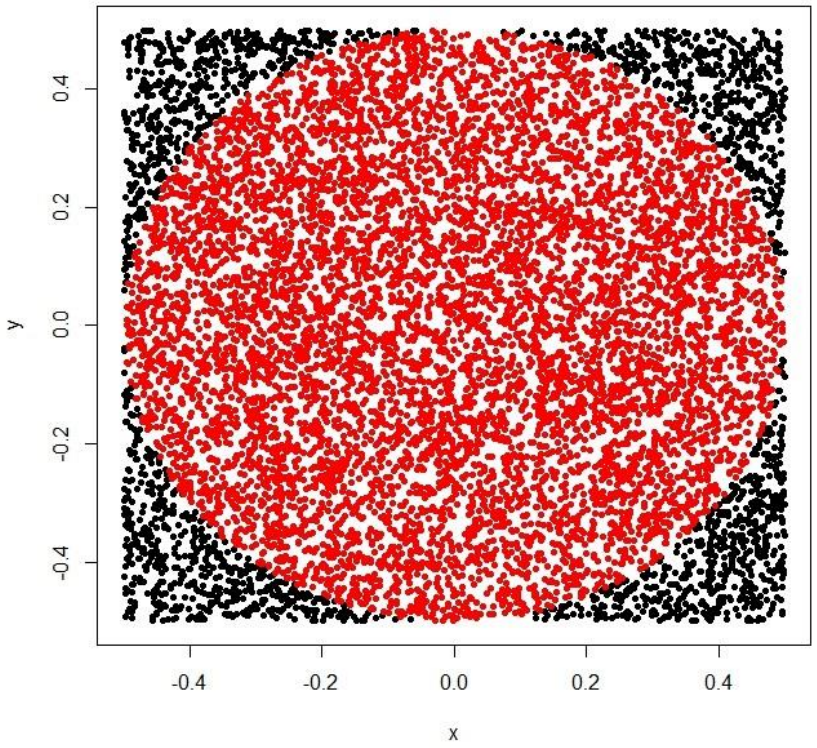
多次元・高次元データ

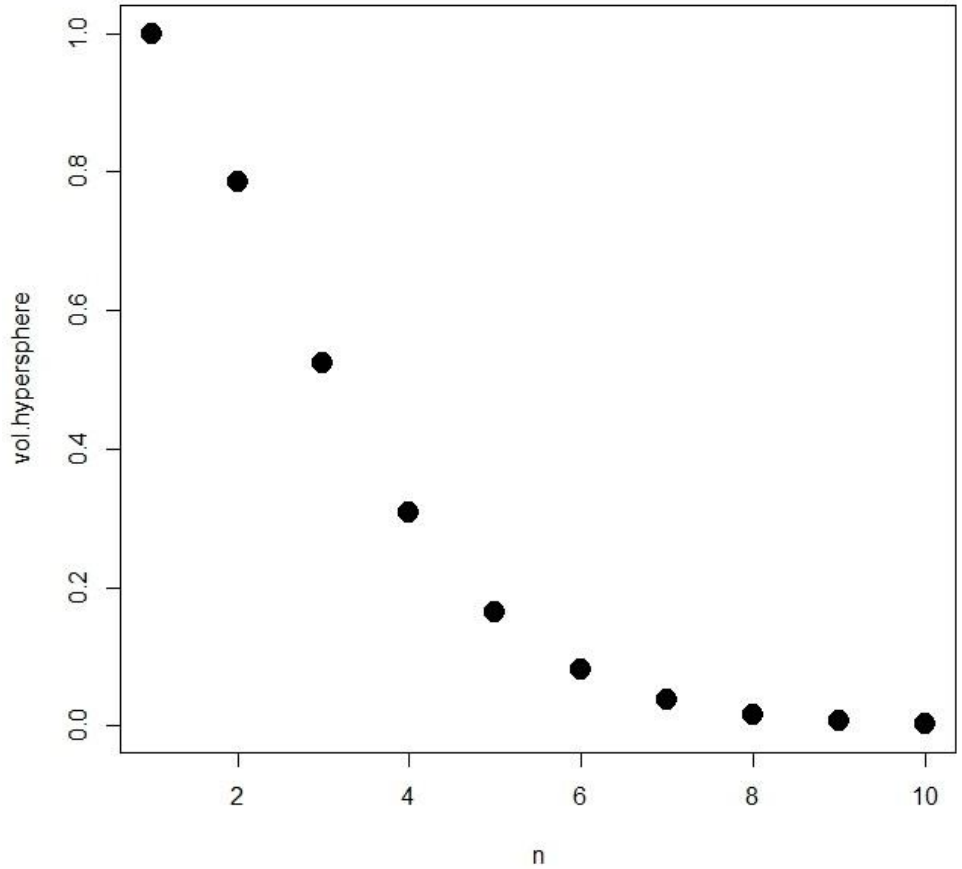
- ２つの大事なこと
  - 「普通」がいない
  - すかすか

多次元・高次元データ

- 「普通」はいない
  - 中央付近: 立方体の中にある球

3.14 / 4 = 0.785





多次元・高次元データ

- Sparse　疎
- 密度を計算するには、単位体積あたりのサンプル数が、そこそこないとうまく行かない。

  - Dim = 1 : 0.1
  - Dim = 2 : 0.01
  - Dim = 3 : 0.001
  - ….
  - Dime = 6 : 0.000001

❶ Density estimation in the high-dim space

"Curse of dimensionality"

As the dimension increases,
the # of bins to consider increases.

dim = 1　　dim = 2　　dim = 3

When # of bins > # of samples, simple "histograms" do not work

$$P(\boldsymbol{x}) = \frac{\#\text{ of points in the same bin}}{\text{total }\#\text{ of points}} \cdot \frac{1}{\text{volume of bin}}$$

多次元・高次元データ

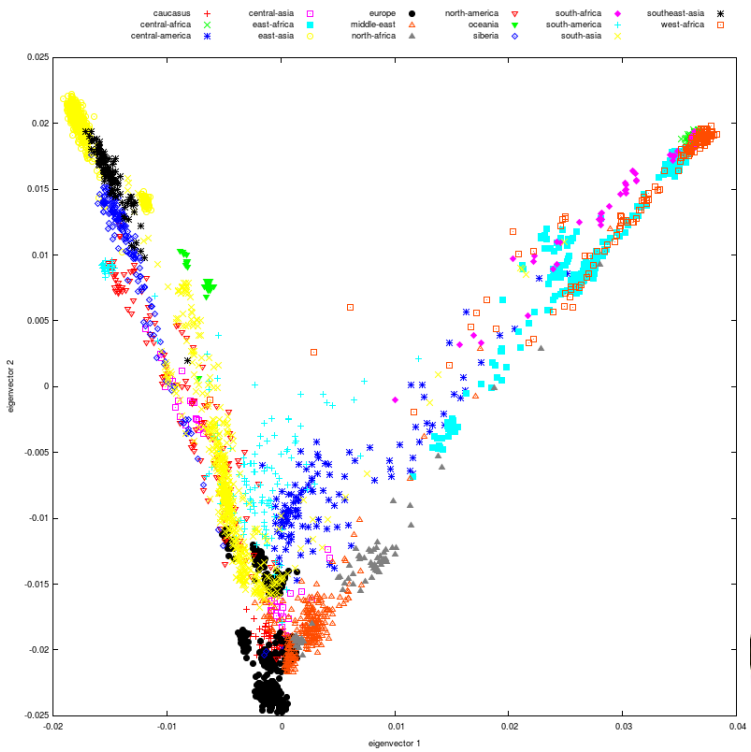
- 広すぎる空間、それなりに「密度」がある

- 高次元空間に低次元多様体として存在している

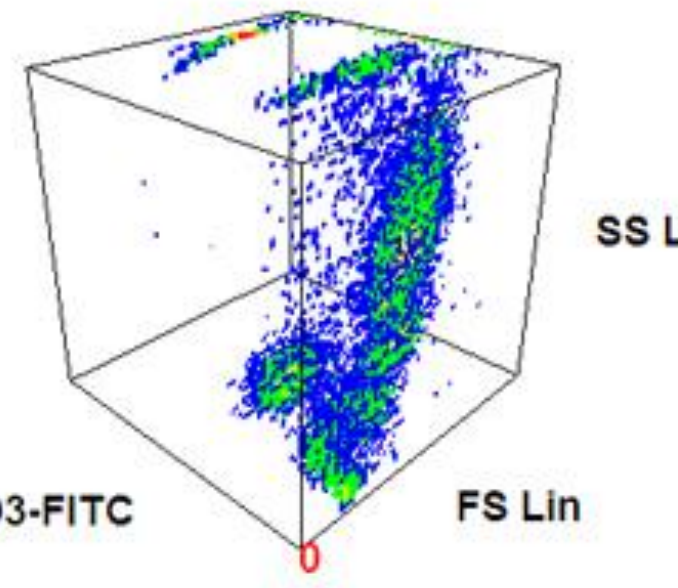
多次元・高次元データ

- ライフサイエンスデータは、高次元空間データとして観察されるが
  - 観察項目が膨大だから
- 項目間の類似・制約も大きく、思ったよりも低次元と、思われている

Ethnic diversity

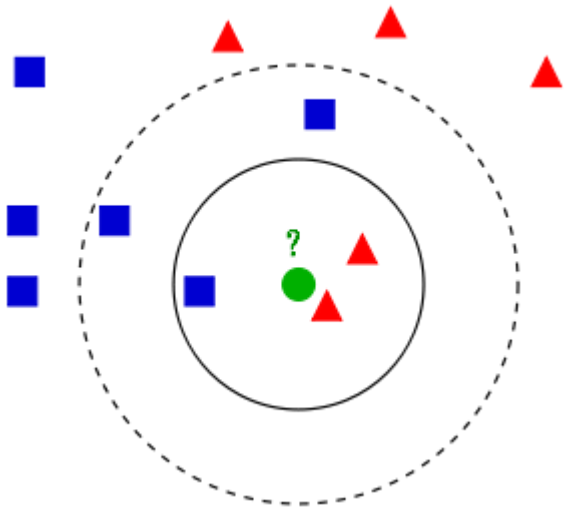


FACS



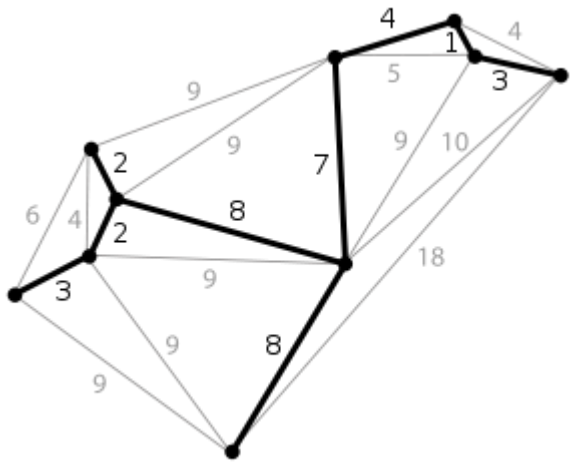
高次元空間にある低次元多様体
その局所密度

- 普通の方法では密度の計算がうまく行かない
- 狭い範囲に区切っても、高次元だと広々しているから



- 密度計算にも工夫
  - K近傍法(k-nearest neighbor法)

- グラフでも似た発想
  - 最小全域木
  - 「近いかどうかだけ」はわかる



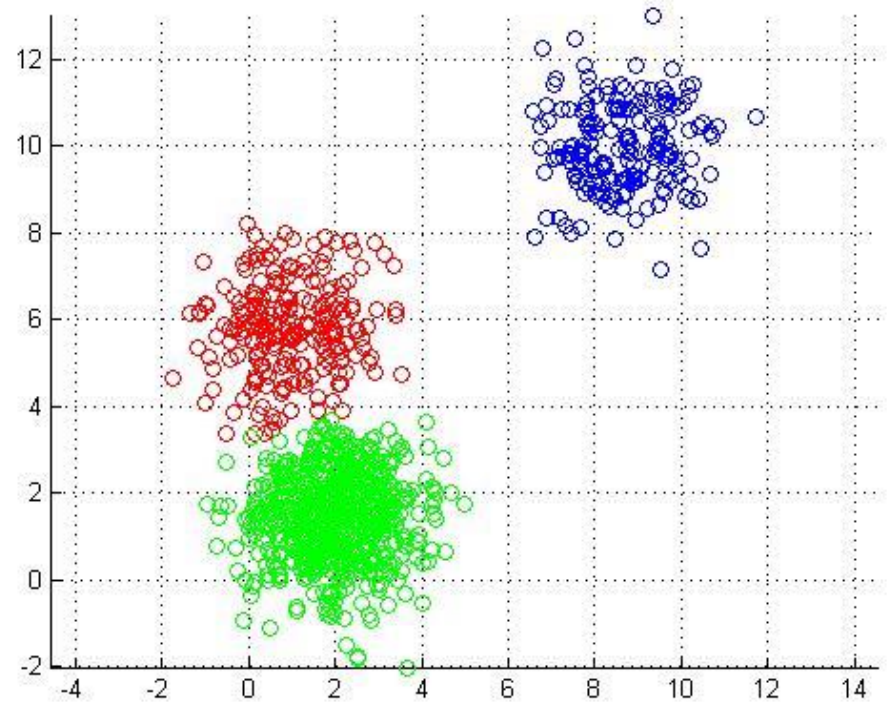
高次元だけれど、思ったより、すかすか

高次元だけれど、思ったより、すかすか

- その、すかすかな加減が
- １次元多様体の点在
- ただし、ばらつきが大きい

高次元だけれど、思ったより、すかすか

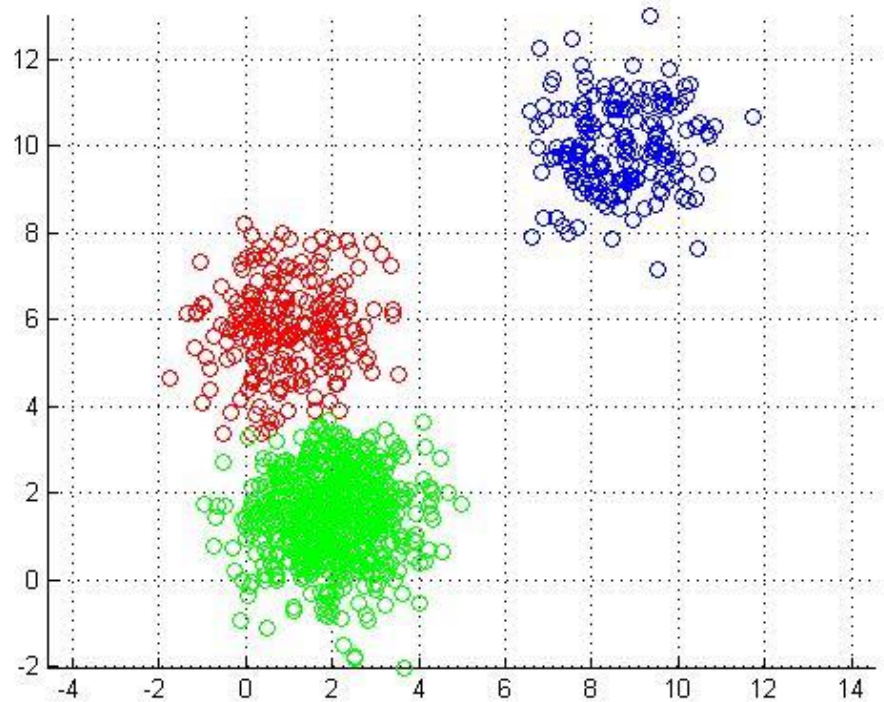
- その、すかすかな加減が
- １次元多様体の点在
- ただし、ばらつきが大きい



高次元だけれど、思ったより、すかすか

- その、すかすかな加減が
- １次元多様体の点在
- ただし、ばらつきが大きい

クラスタリング

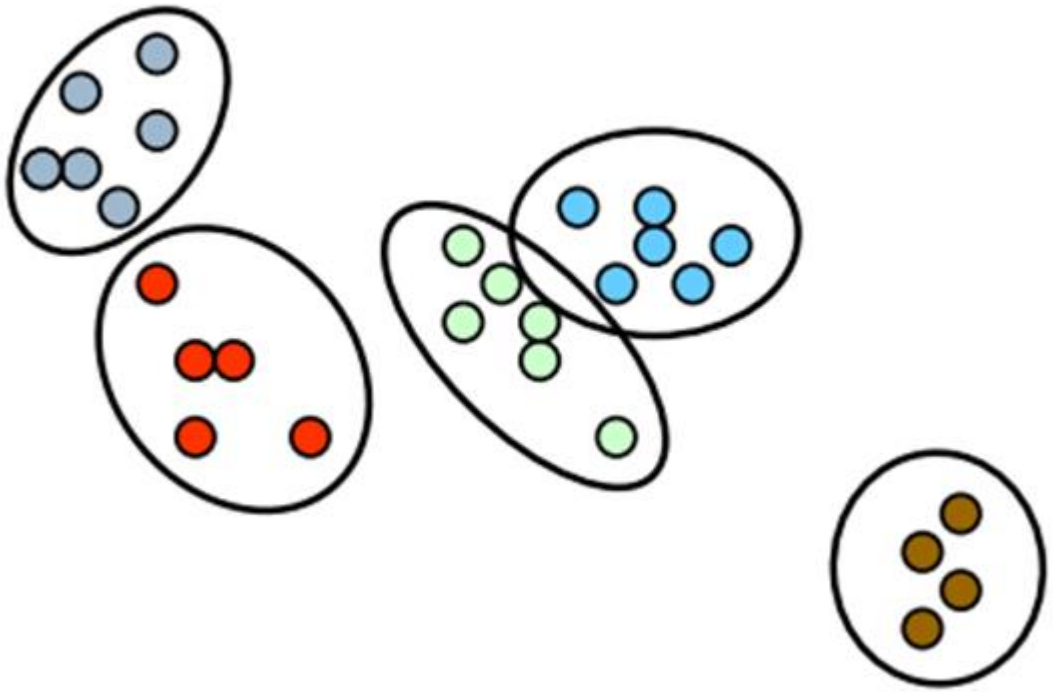


検定・推定・分類

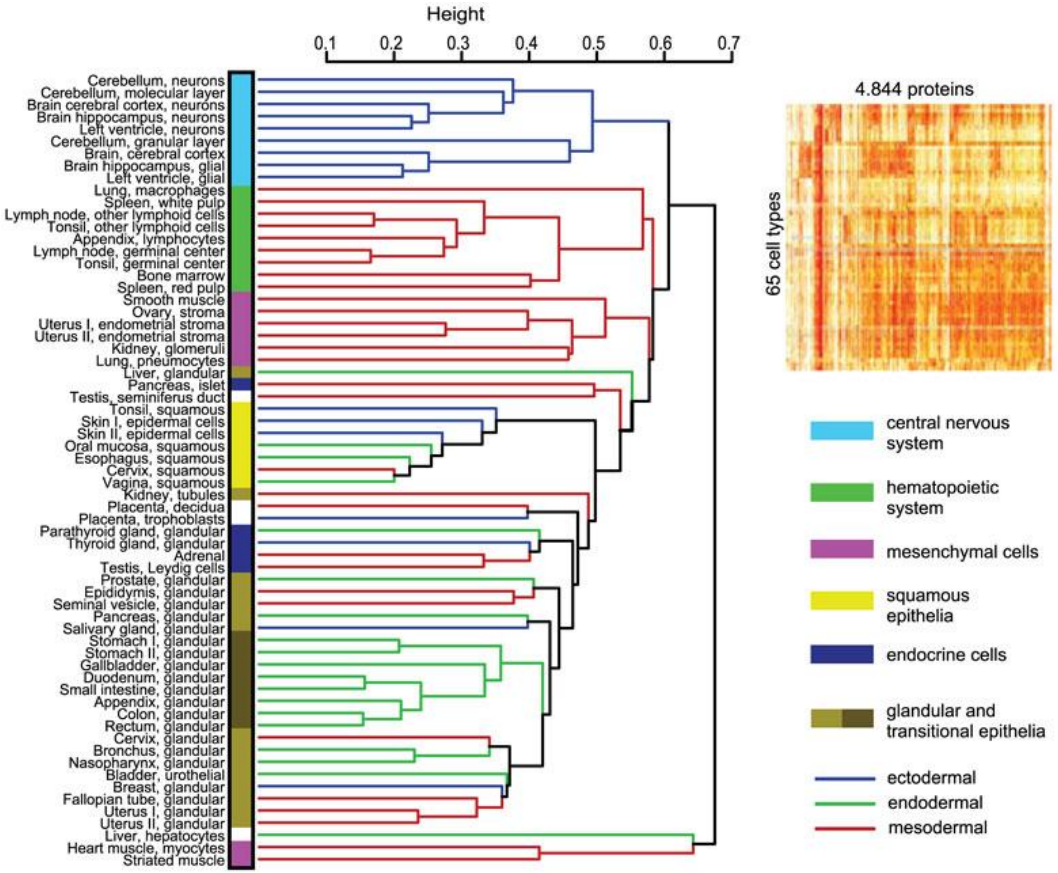
- 検定
  - 有意、エラーのコントロール、多重検定
- 推定
  - 区間推定、モデル推定、ベイズ
- 分類
  - 教師アリ、教師ナシ

クラスタリングの方法、2タイプ

・非階層的

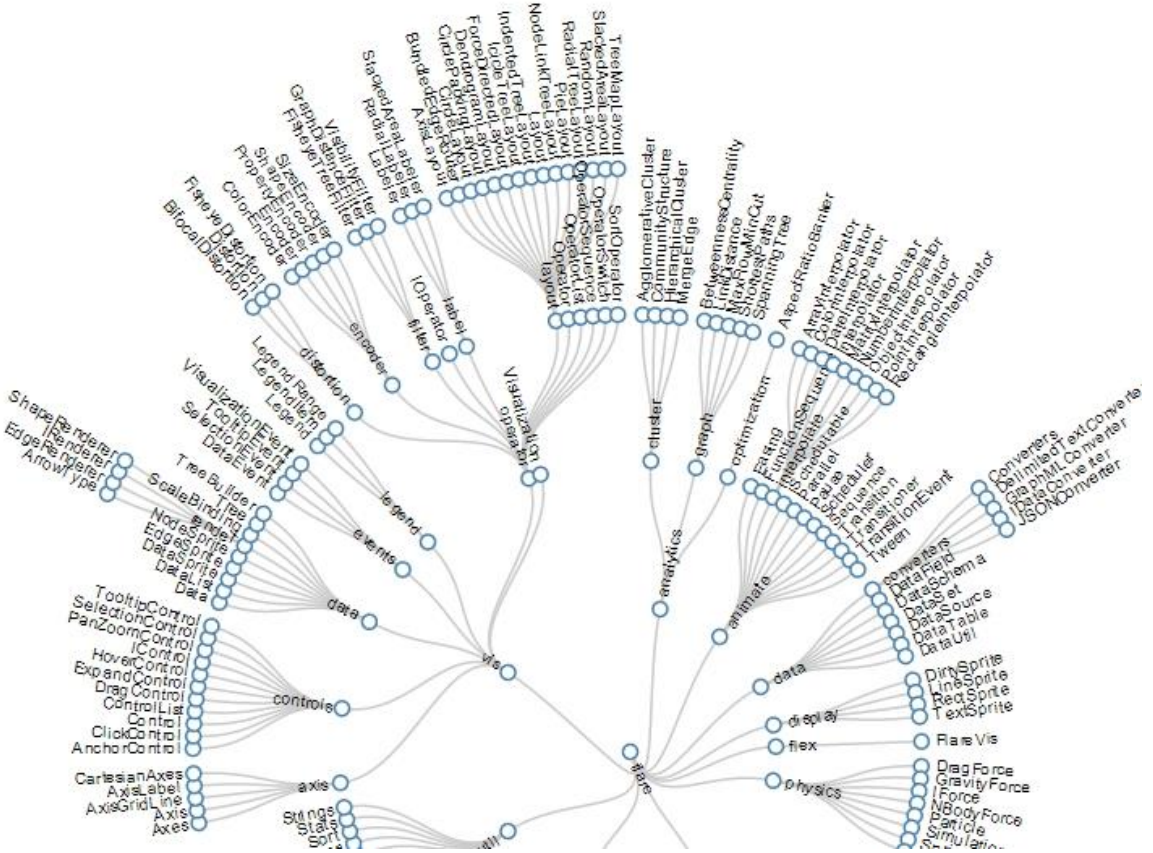


・階層的



階層的

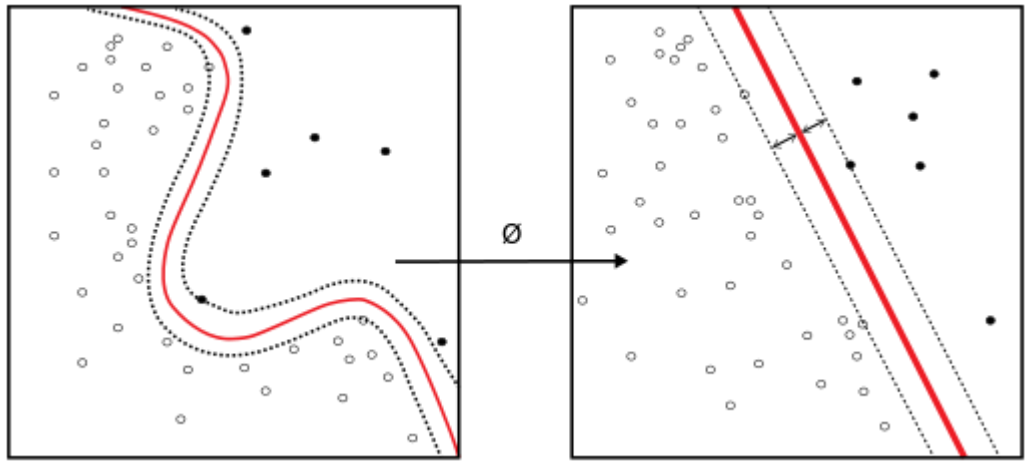
- 木構造　～　グラフ、再び
- 構造が持つ情報
- 構造が持つ次元
- 木の上に定義された距離

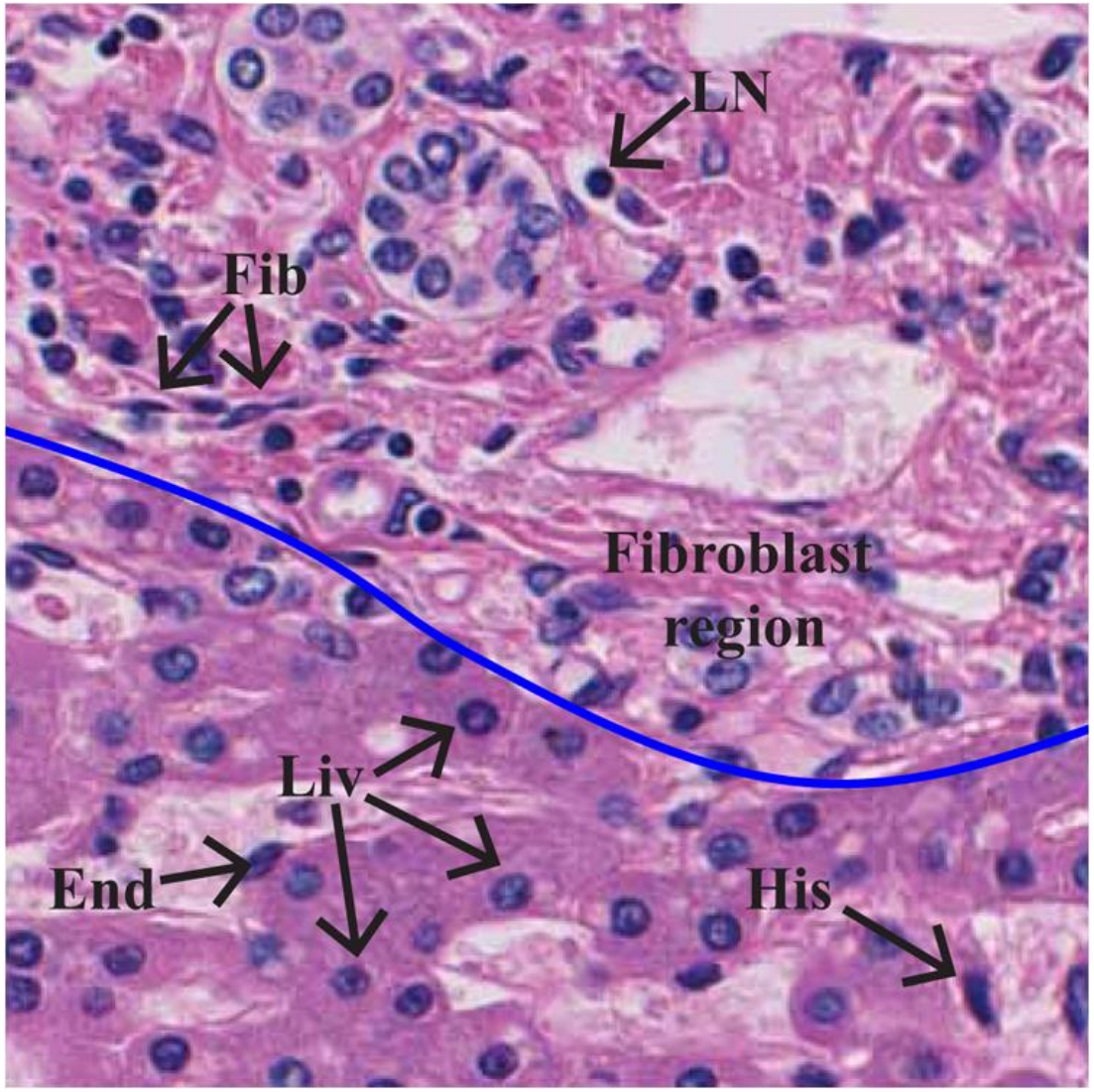


- そもそも、木構造をとるべき対象には階層的クラスタリングが向いている

分類

- 分けにくい広がりを分ける

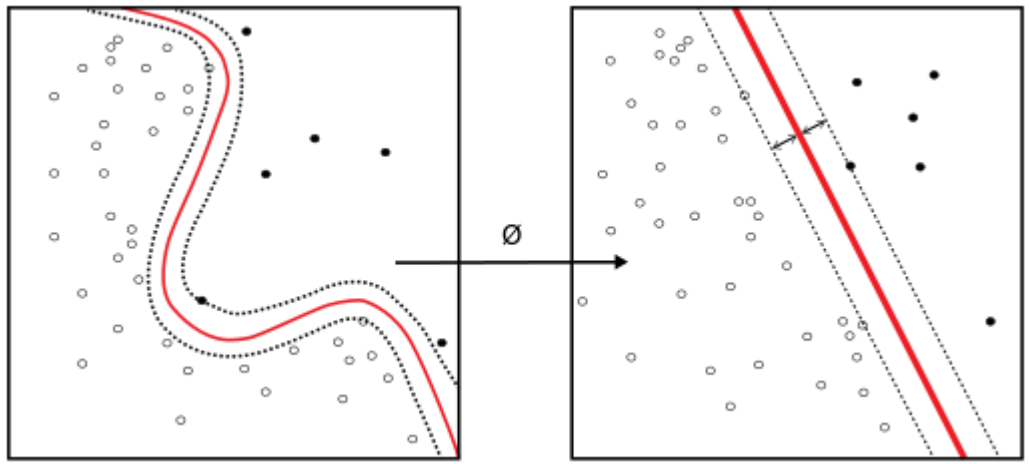




J. Med. Imag. 1(3), 034501 (Oct 09, 2014). doi:10.1117/1.JMI.1.3.034501

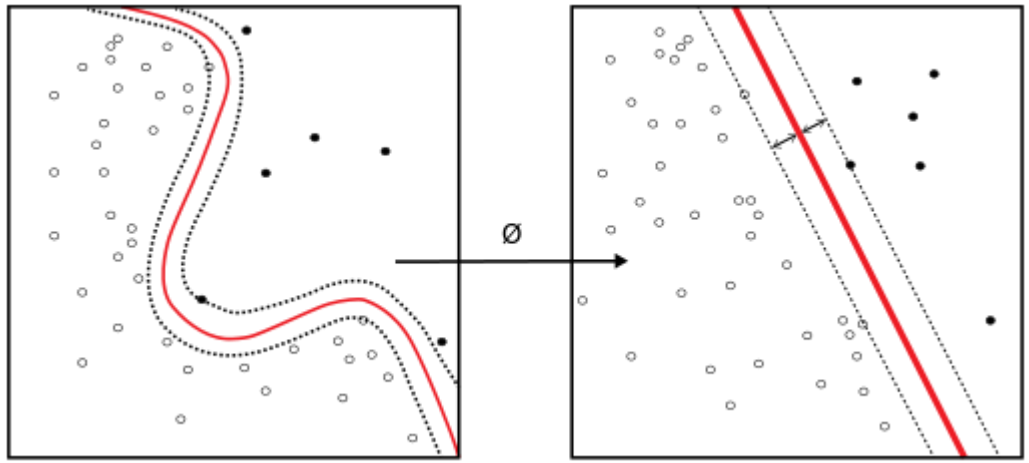
分類

- 教師なし学習
- 教師あり学習



分類

- 教師なし学習
- 教師あり学習
- 答えはないけれど、「当たる方法かどうか」を知りたい
  - クロス-バリデーション：リサンプリング法



ゲノム・オミクス研究における統計・データサイエンスの役割

- ノイズのあるハイスループットデータのデータQC
- 検定・推定・分類
- 多次元・高次元データ
- 乱数を使ったアプローチ
- その他：実験デザイン

ゲノム・オミクス研究における統計・データサイエンスの役割

- ノイズのあるハイスループットデータのデータQC
- 検定・推定・分類
- **多次元・高次元データ**
- 乱数を使ったアプローチ
- その他：実験デザイン

まだ、もう少し高次元問題を

Small n Large p

- サンプルサイズ 100
  - ある一つの遺伝子の発現量とある表現型との関係を検定する
  - N = 100, p = 1
  - Large n Small p
- サンプルサイズ 100
  - たくさんの遺伝子の発現量とある表現型との関係を検定する
  - N = 100, p = 25000
  - Small n Large p

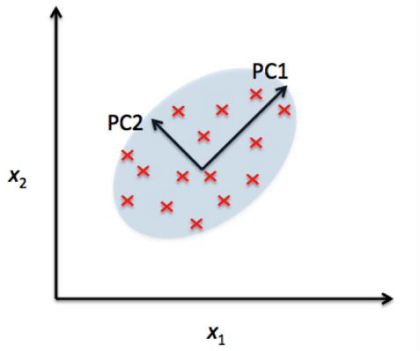
n = p は解ける。完璧な回帰

- q = a x; q = 3, x = 2 → 解ける

- q1 = a x1 + b y1
- q2 = a x2 + b y2　→　解ける

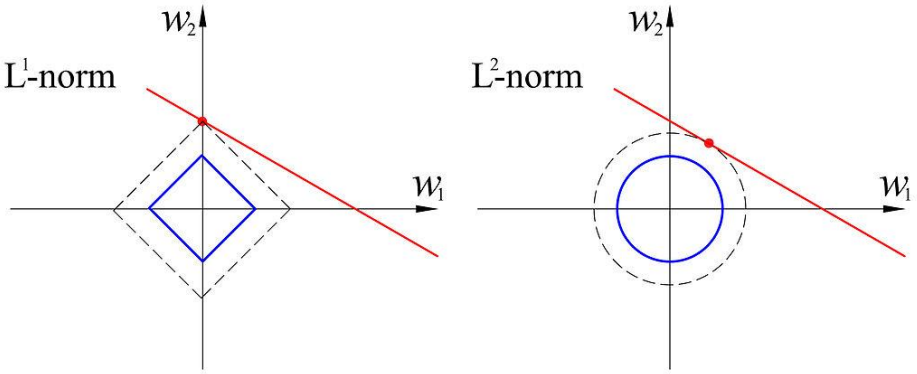
- q1 = a x1 + b y1 + c z1
- q2 = a x2 + b y2 + c z2
- q3 = a x3 + b y3 + c z3 →　解ける

n << p



- ある変数セットで、完璧な回帰ができる
- 別の変数セットでも完璧な回帰ができる

- どのセットがよいかわからない



- 完璧な回帰ができるのがよいわけでもない
  - AIC ～ Simpler model is better
  - LASSO, Sparse

  - k << n 個の変数で説明できるはず…事前予想～ベイズ

ゲノム・オミクス研究における統計・データサイエンスの役割

- ノイズのあるハイスループットデータのデータQC
- 検定・推定・分類
- 多次元・高次元データ
- 乱数を使ったアプローチ
- その他：実験デザイン

ゲノム・オミクス研究における統計・データサイエンスの役割

- ノイズのあるハイスループットデータのデータQC
- 検定・推定・分類
- 多次元・高次元データ
- 乱数を使ったアプローチ　～　モンテカルロ法
- その他：実験デザイン

リサンプリング

- 標本から統計量を推定する
  - ジャックナイフ（サブセット）、ブートストラップ(Replacement)
- 統計的有意差
  - パーミュテーション（順列）～正確確率
- クロス-バリデーション

リサンプリング

- 標本から統計量を推定する
  - ジャックナイフ（サブセット）、ブートストラップ(Replacement)
- 統計的有意差
  - パーミュテーション（順列）～正確確率
- クロス-バリデーション

- 乱数を使う～計算機による疑似乱数列

疑似乱数列

- 一様分布から
- 既存の分布から

疑似乱数列

- 一様分布から
- 既存の分布から

- 任意の分布から Gibbs sampling

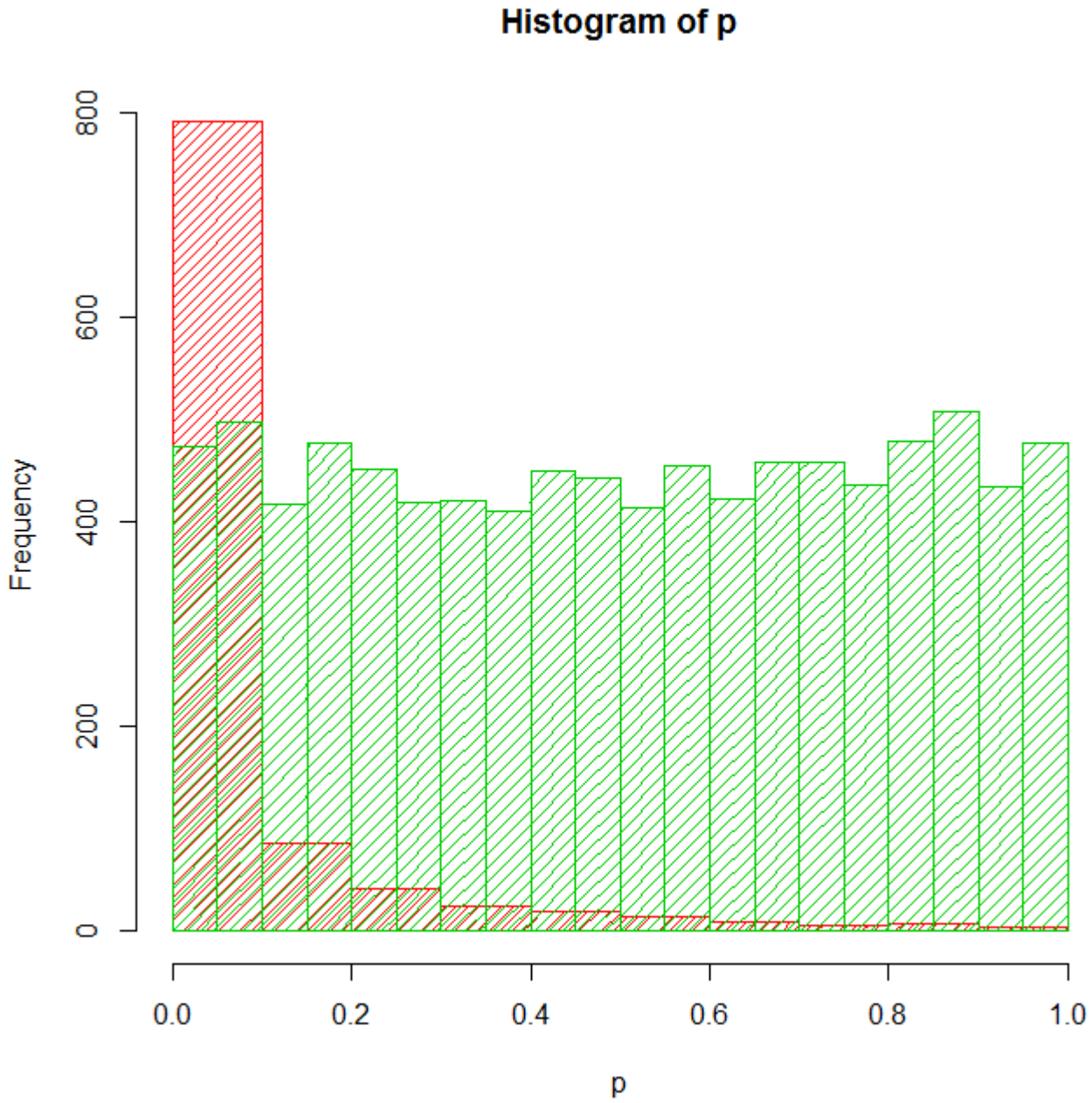
疑似乱数列

- 一様分布から
- 既存の分布から

- 任意の分布から Gibbs sampling
- Gibbs sampling を利用して
  - 確率モデルを構成して、その確率分布のパラメタを推定しながら、その推定分布から乱数を発生させて…
  - BUGS (Bayesian inference using Gibbs Sampling)

例

- 赤と緑の比率を推定しながら
- 赤の分布を非心カイ二乗分布と仮定しつつ、その非心パラメタを推定しながら
- 「比率」と「非心パラメタ」との両方との最尤推定値を、モンテカルロ法で推定する



疑似乱数列

- 一様分布から
- 既存の分布から

- 任意の分布から Gibbs sampling
- Gibbs sampling を利用して
  - 確率モデルを構成して、その確率分布を推定しながら
  - BUGS (Bayesian inference using Gibbs Sampling)
- MCMC(マルコフ連鎖モンテカルロ）でシミュレーション
  - それにStan (ベイズ推定ソフトウェア）をかぶせる

疑似乱数列・モンテカルロ

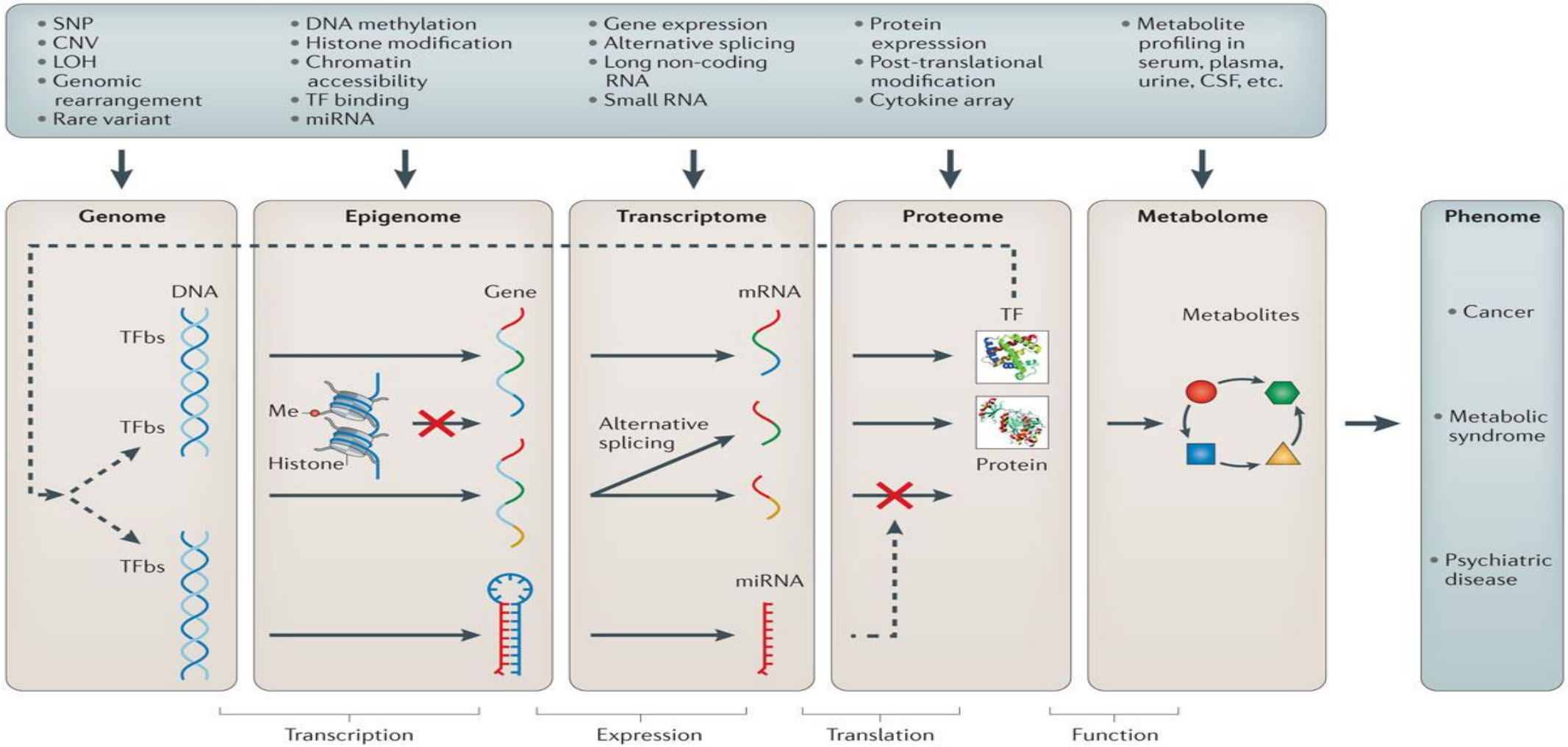
- コンピュータ・エイジの手法

ゲノム・オミクス研究における統計・データサイエンスの役割

- ノイズのあるハイスループットデータのデータQC
- 検定・推定・分類
- 多次元・高次元データ
- 乱数を使ったアプローチ
- その他：実験デザイン

実験デザイン

- さまざまなデータ
- 全部合わせて、何を言う？



Nature Reviews | Genetics

個別も大変、合わせるのはもっと大変

- モデル・合わせるための仮定 を立てて合わせる
  - 合わせ方の構造も色々なやり方がある
  - データ自体を統合して使う
  - 個々の解析の結果を統合する（いわゆるメタ解析）
    - 同じフォーマットからの結果の統合が本来のメタ解析
- 個々の解析アプローチに違いが大きいので合わせにくい
  - 解析アプローチ固有の要素を排除して、個別解析自体を「統合しやすいもの」に置き換える

資料など

- 本講義のスライドを含め、関連知識・関連資料等が
- http://statgenet-kyotouniv.wikidot.com/statistical-analysis-for-genome-based-life-science からアクセスできます