区間推定_尤度比

法数学勉強会

2016/06/15

京都大学(医)統計遺伝学分野

山田　亮

尤度比

- 尤度比が 3.2×10^7
- あくまでも　推定値
- 真の値は、それより高いかもしれないし、低いかもしれない

推定

- 点推定値

- 区間推定値

推定

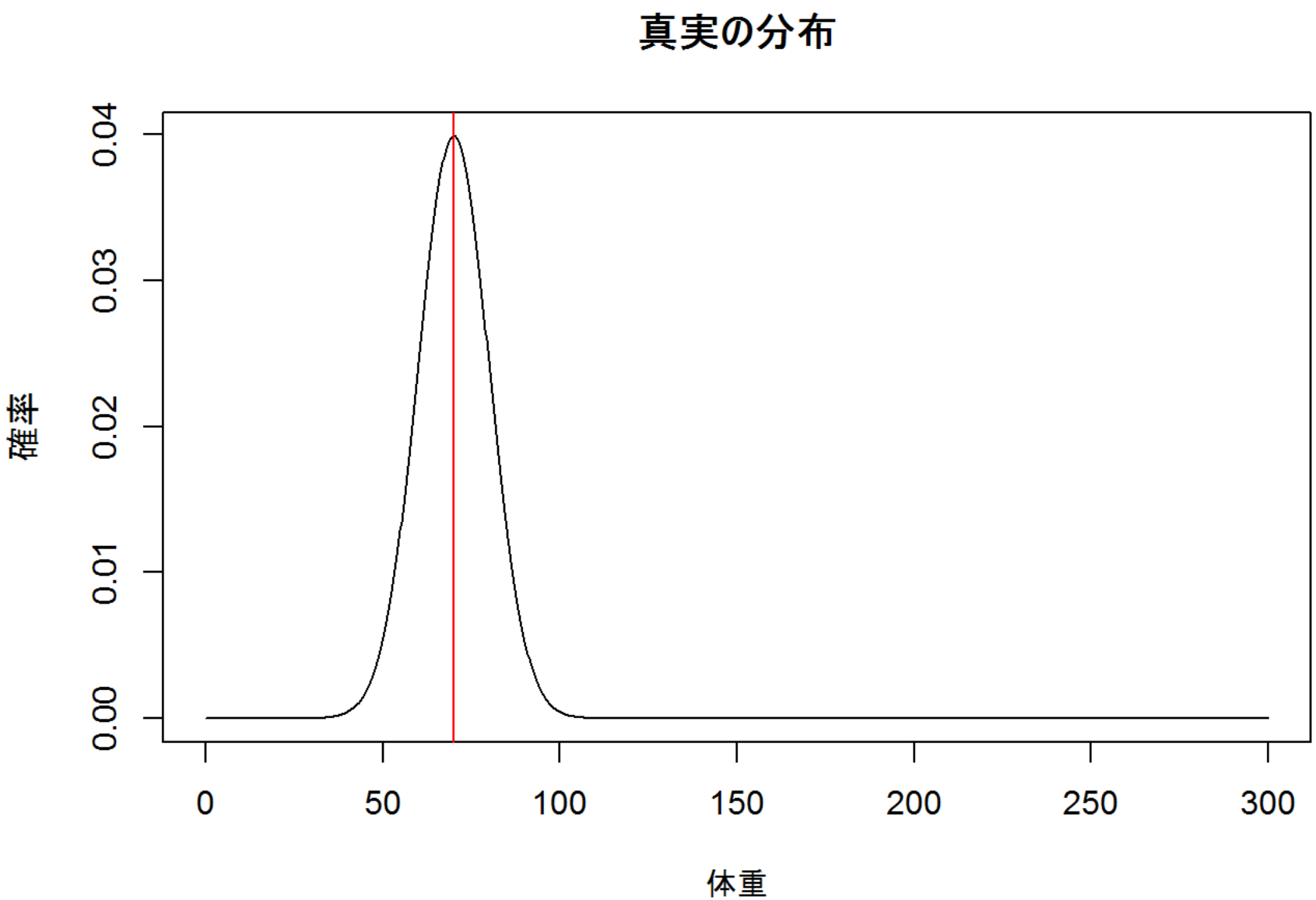
- 点推定値
  - 期待値～平均値
  - 最尤推定値
- 区間推定値

推定

- 点推定値
  - 期待値～平均値
  - 最尤推定値
- 区間推定値
  - 頻度主義信頼区間
  - ベイジアン区間
  - その他いろいろ…

平均体重を推定する

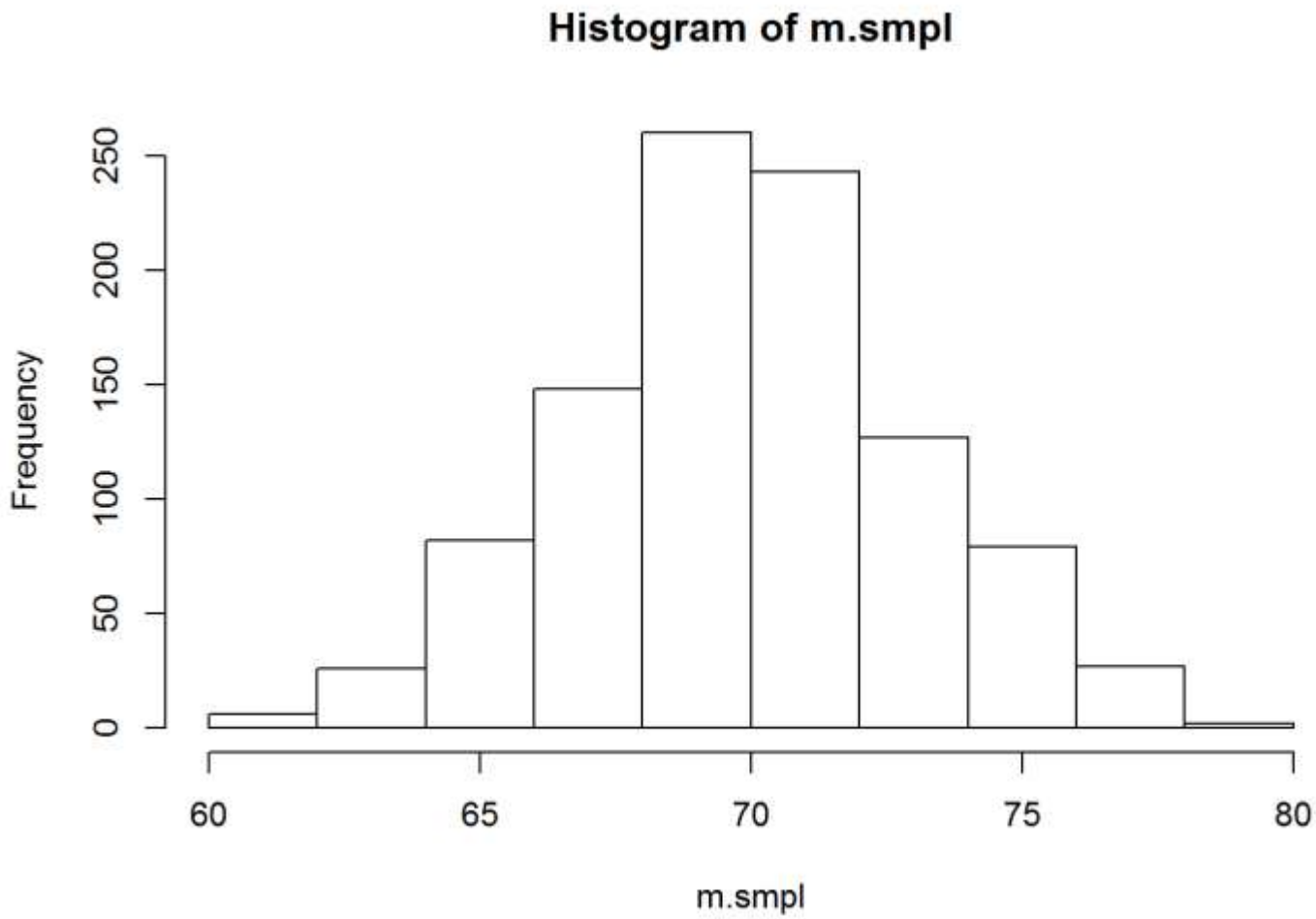
- サンプルの平均値
- 「真実の分布」が平均50、標準偏差10のとき、どうしたら「真実の平均」を知ることができるか？
- 一部のサンプルを取り出して、そのサンプルの平均を計算して、代用する。


真実の分布
確率
0.00
0.01
0.02
0.03
0.04
0
50
100
150
200
250
300
体重

サンプル数10、そのサンプル平均値

```
##  65.36669
##  73.16953
##  68.45564
##  69.41692
##  71.3132
##  73.16633
##  66.35471
##  72.59056
##  70.57345
##  67.77196
```

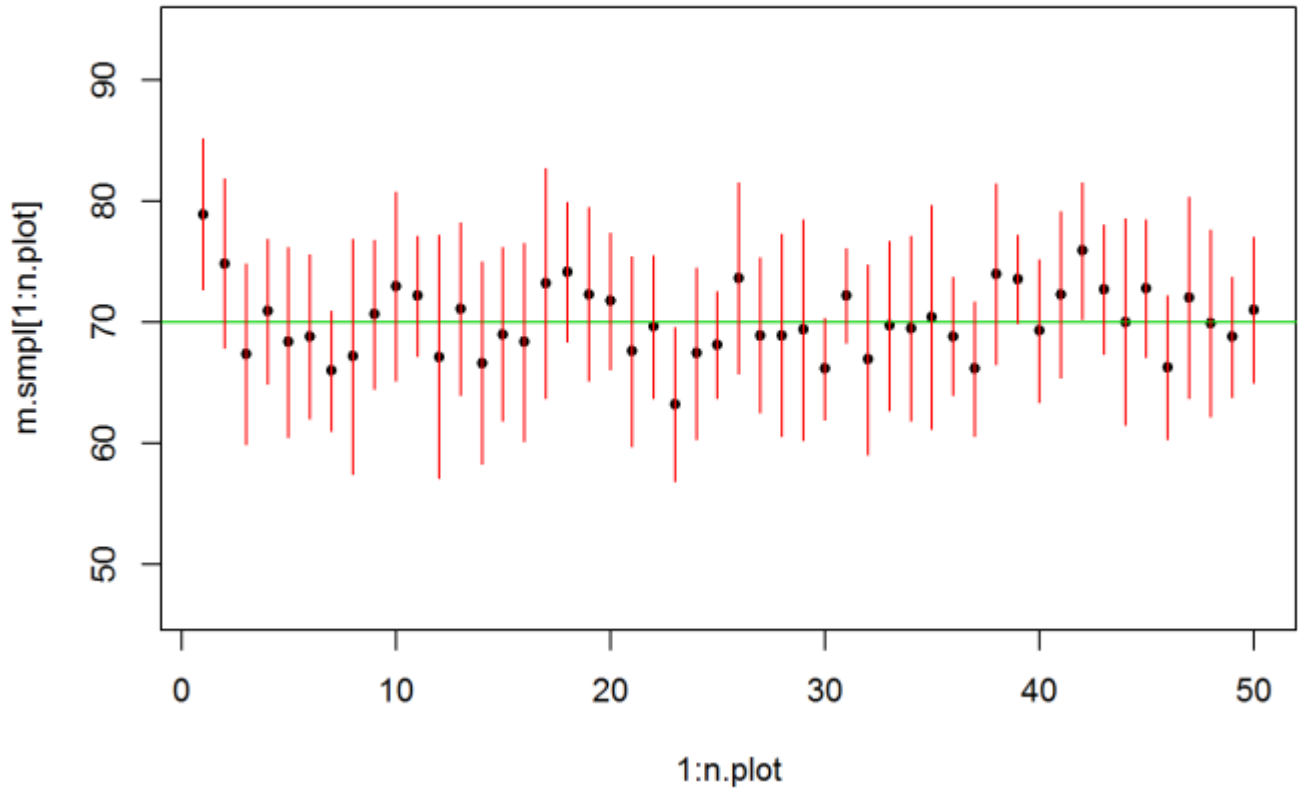
10サンプル、1000回



サンプルから信頼区間

- 本当の値を当てることはできない
- 「ここから、ここの間に真の平均は入る」と言えば、当たる確率が出せる
- 95% 信頼区間とは、
  - 「サンプルがあったときに、その値を使って、『ここからここまでと予想する』というルールを決める」
  - 「そのルールに従うと、95%の場合、真の値が、その範囲に入る」
  - と言うようにデザインされた『ルール』のこと。もしくは、その『ルール』に従って算出した『区間』のこと。

「あるルール」　＝　赤い線
たしかに、1000 回のうち951回は、赤い線が緑を含んだ



- 正規分布を仮定して、比較的簡単に、+ - x / で計算している。
- 一応、式を載せますが、今日は、式は気にしないで行きます。

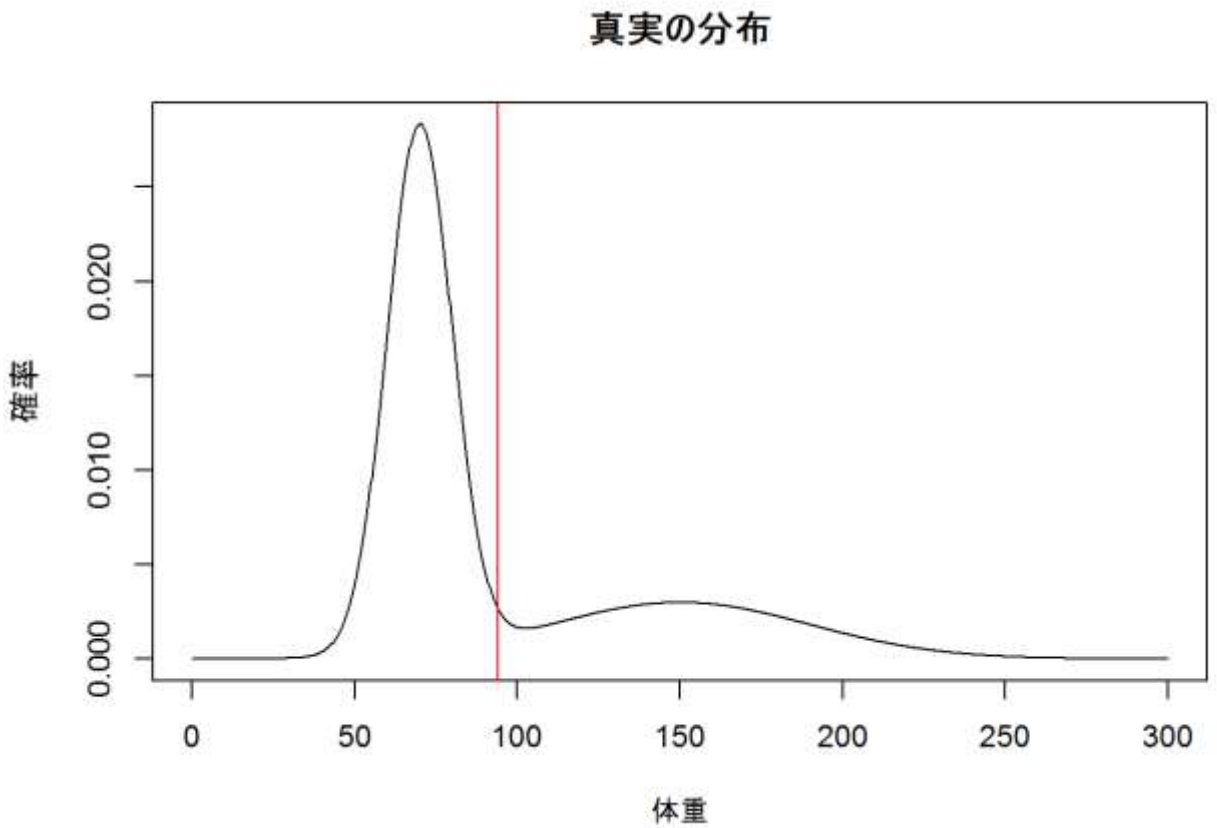
$$m \pm k\sqrt{\frac{a}{n}}$$

$$m = \frac{\sum x_i}{n}$$

$$a = \frac{\sum (x_i - m)^2}{n - 1}$$

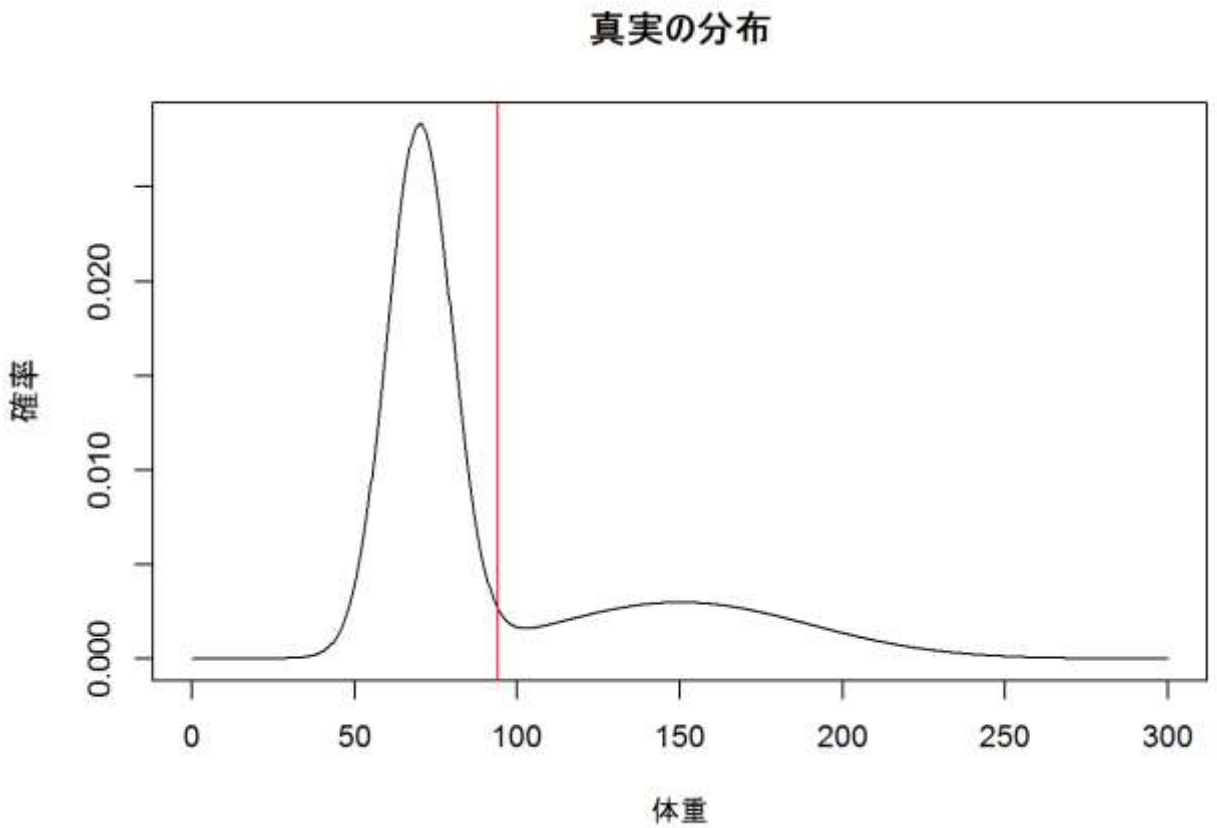
分布がきれいでないとき

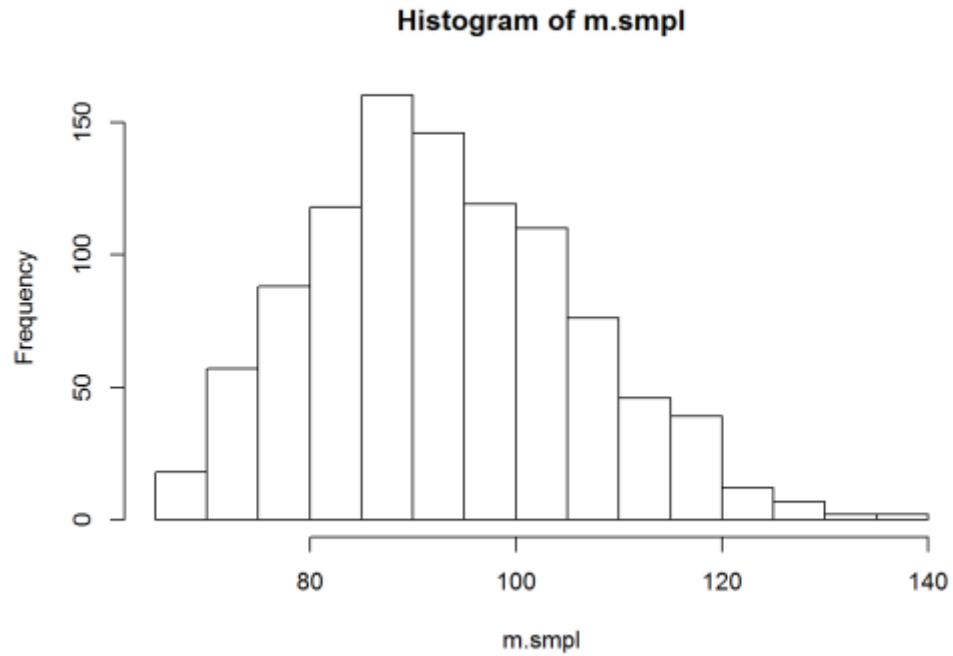
- 正規分布でないとどうなるか。



分布がきれいでないとき

- 正規分布でないとどうなるか。

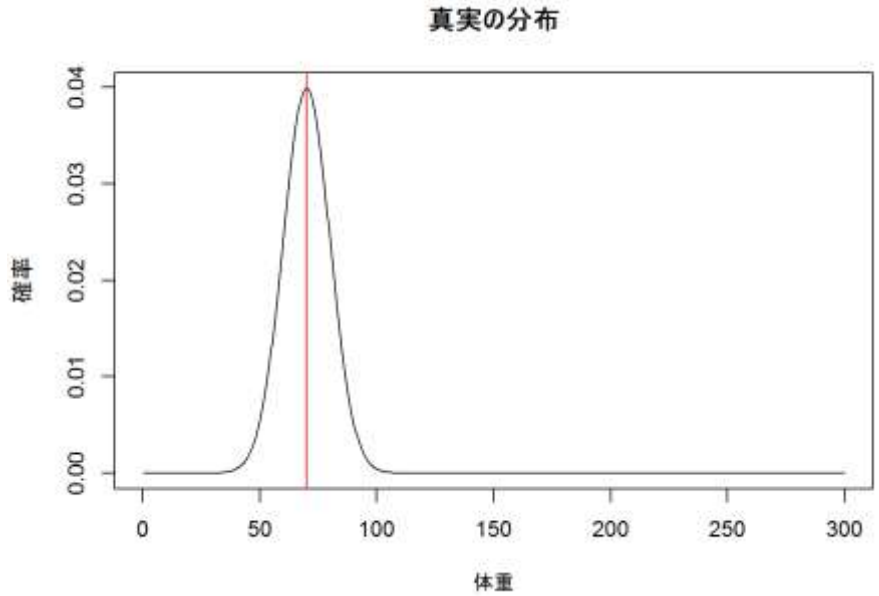


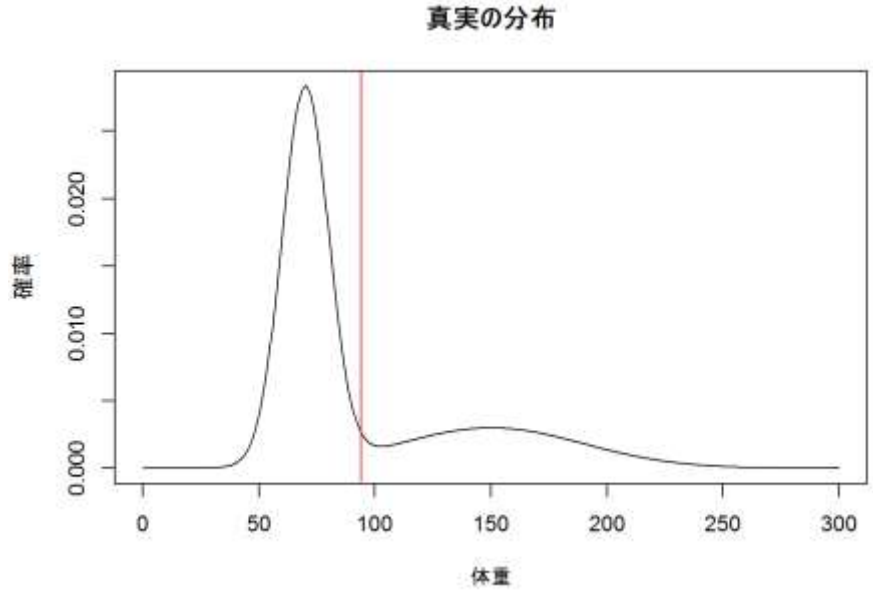


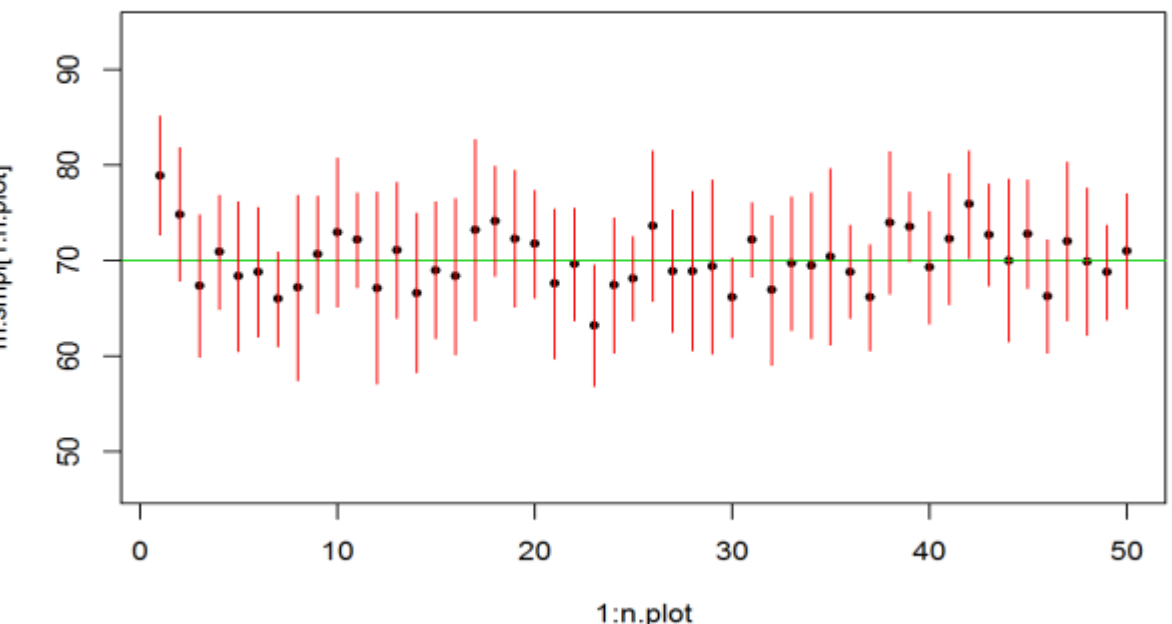
「信頼区間」はあたっているのか？

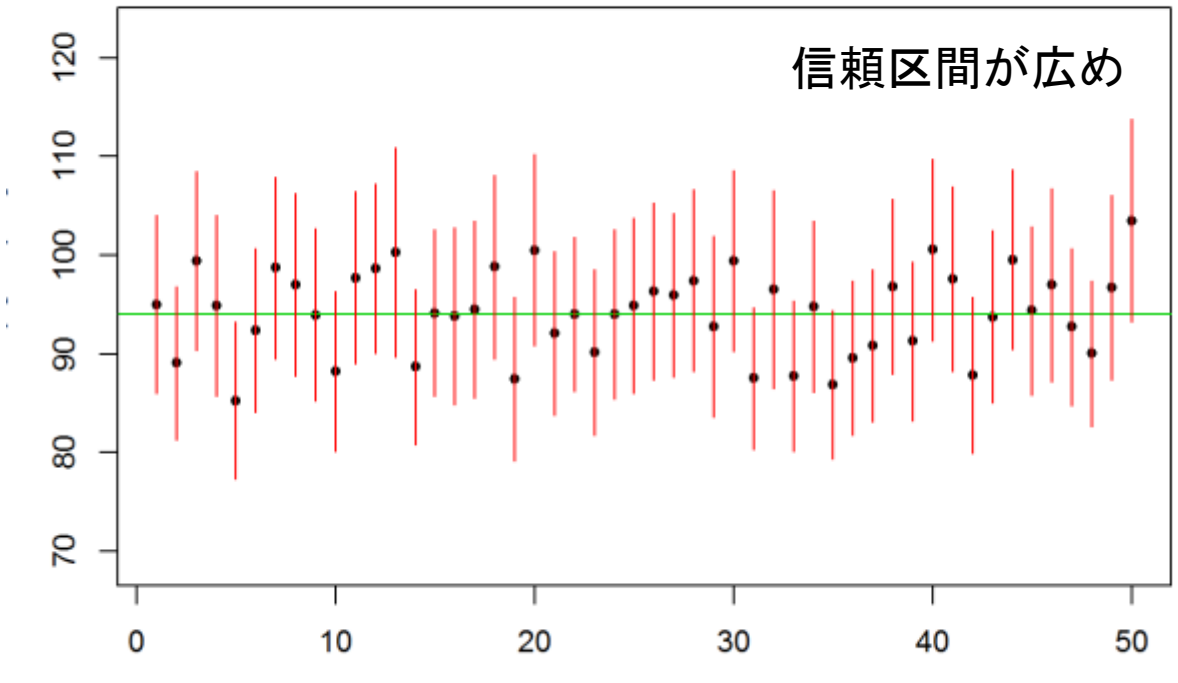
- サンプル数１０
  - 当たった確率　0.887
- サンプル数を増やせば (サンプル数　100)
  - 当たった確率　0.940

- 真の分布をだいたいカバーできれば、当たる。真の分布の複雑さに比べてサンプル数が少なすぎると、当たらなくなる。


真実の分布
確率
0.00
0.01
0.02
0.03
0.04
0
50
100
150
200
250
300
体重


真実の分布
確率
0.000
0.010
0.020
0
50
100
150
200
250
300
体重


50
60
70
80
90
0
10
20
30
40
50
1:n.plot


信頼区間が広め
70
80
90
100
110
120
0
10
20
30
40
50

DNA鑑定のための区間推定

- 体重の区間推定がしたいわけではない。
- DNA型ジェノタイプが、たまたま一致する尤度を計算するためには、ジェノタイプ頻度を推定したい。

頻度推定

- 簡単のために、「あたり vs. はずれ」という枠組みで、成功率を推定することにする。
- 確率pで当たりが出るくじ引きがある。
- １０回引いて、３回当たった。
- さて、pはいくつか？
- その信頼区間は？

成功＝１、失敗＝０

- 真の成功率は0.05
- ３０回の試行、１回の成功
- 「成功率」を「成功と失敗の平均」と考えれば、体重のときと同じことができる。平均成功率とその信頼区間とみなせば…

- 平均　0.033333

成功＝１、失敗＝０

- 真の成功率は0.05
- ３０回の試行、１回の成功
- 「成功率」を「成功と失敗の平均」と考えれば、体重のときと同じことができる。平均成功率とその信頼区間とみなせば…

- 平均　0.033333
- ー0.03484099 ～ 0.10150765

成功＝1、失敗＝0

- 真の成功率は0.05
- 30回の試行、1回の成功
- 「成功率」を「成功と失敗の平均」と考えれば、体重のときと同じことができる。平均成功率とその信頼区間とみなせば…

- 平均　0.033333
- −0.03484099 ～ 0.10150765

- マイナス！

区間推定をするときには考慮するべきことがある

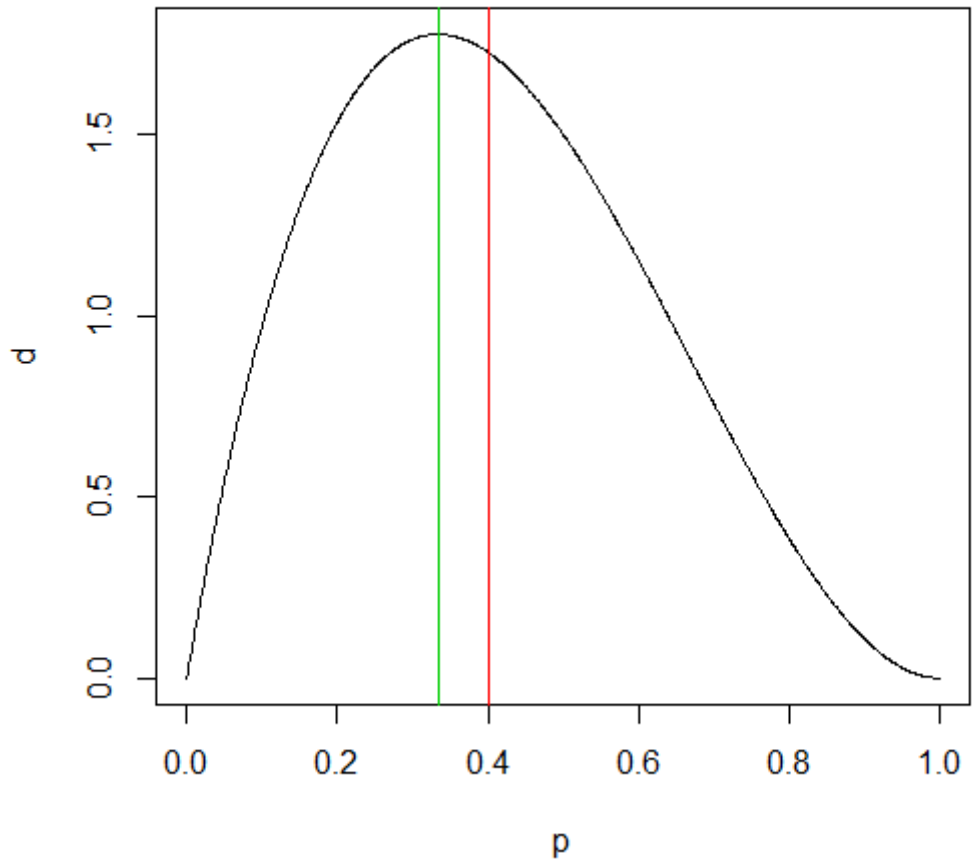
- 信頼区間に「負」があるのはどうして『いけない』か？
- 成功率は0から1だと「知っている」から。

ベイズ推定

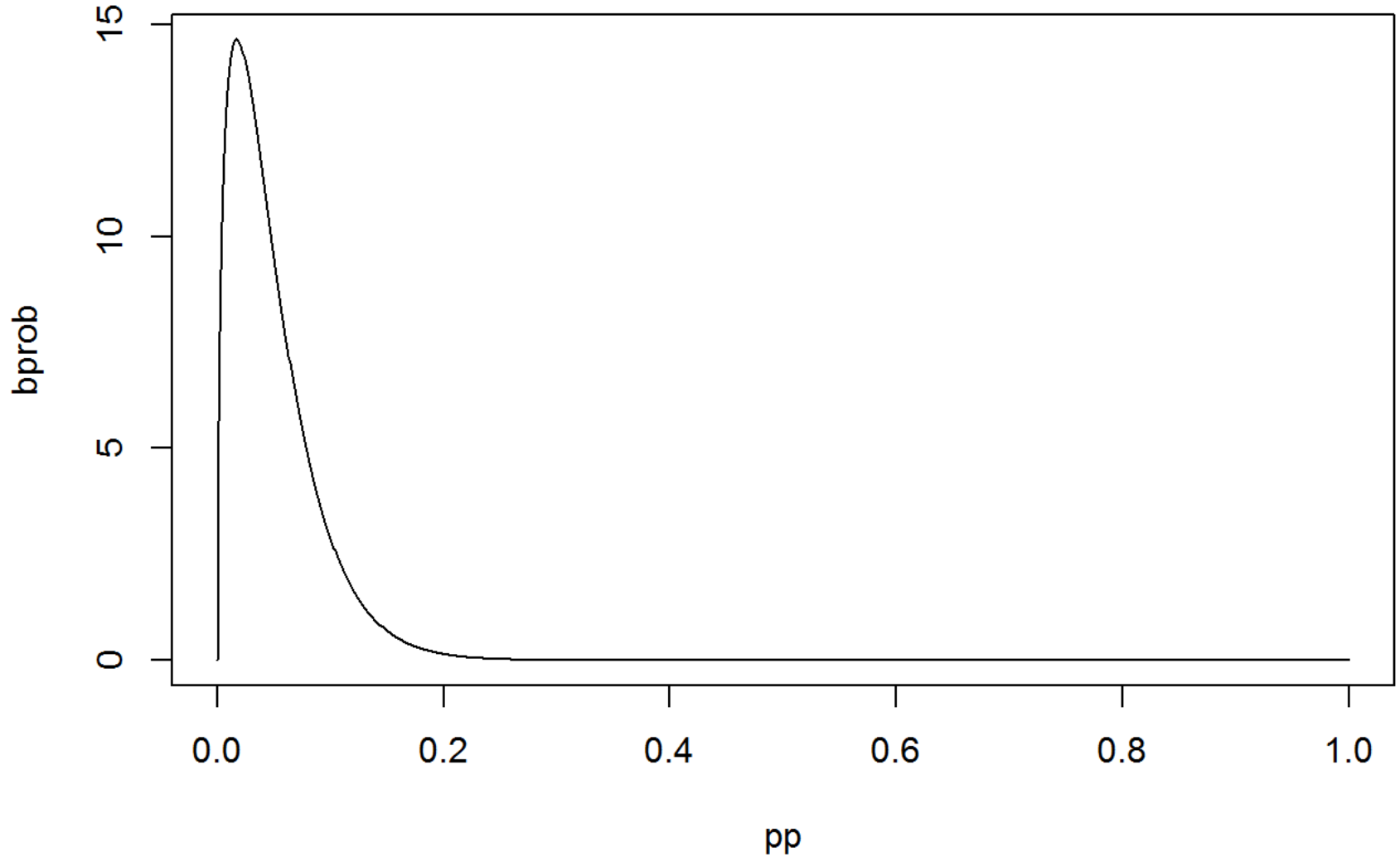
- DNA鑑定界ではベイズ推定の方が主流
- よく考えたら、二項分布の観察はベータ分布でベイズ推定もできたはず…

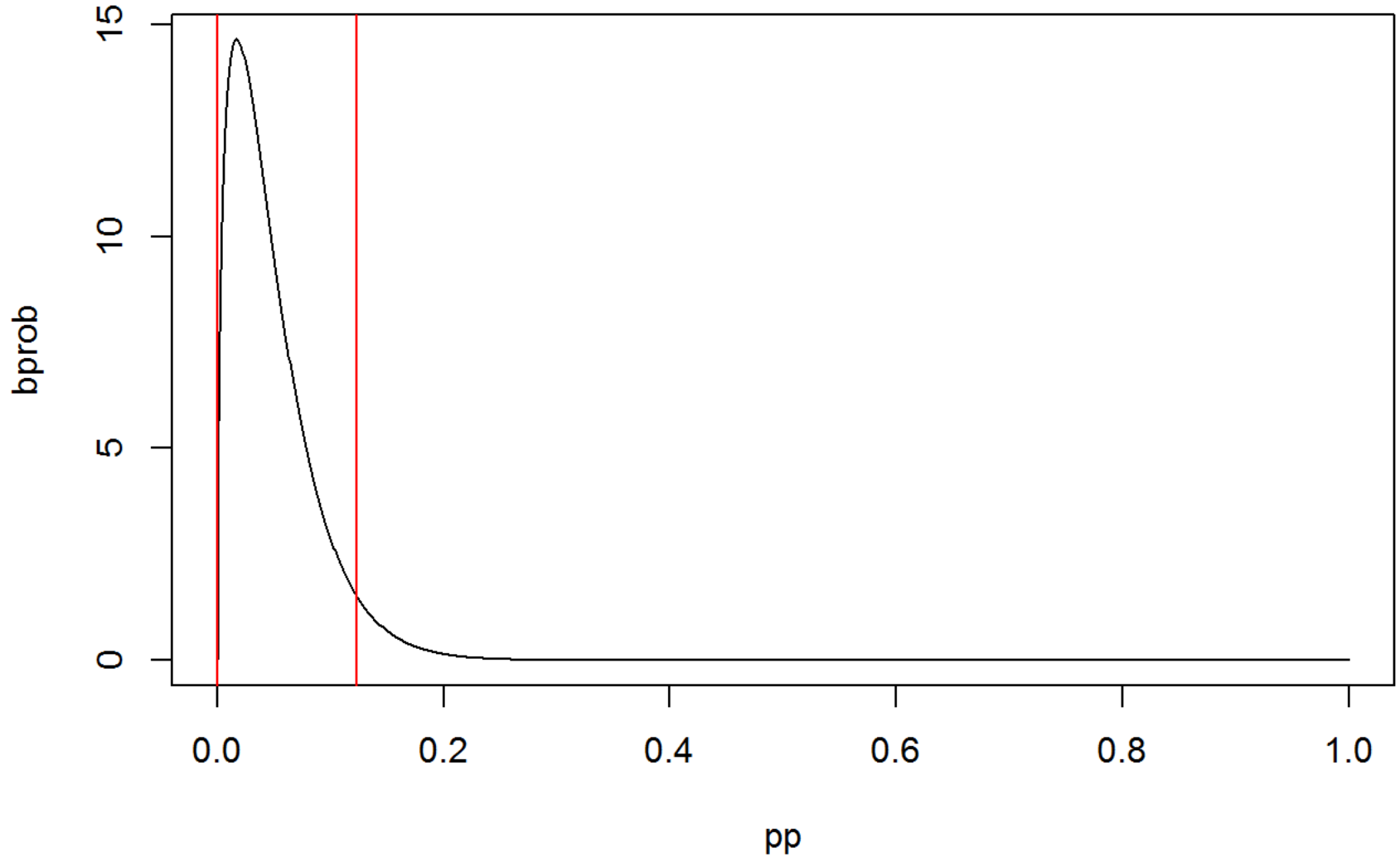
3回引いて、1回の当たり

最尤推定値　1/3



期待値　(1+1)/(3+2)


bprob
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
pp


bprob
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
pp

区間推定はベイズ推定だけではない

```
##              method x  n       mean          lower      upper
## 1  agresti-coull 1 30 0.03333333 -8.305484e-03 0.18091798
## 2     asymptotic 1 30 0.03333333 -3.090070e-02 0.09756737
## 3         bayes 1 30 0.04838710  6.903016e-05 0.12314380
## 4        cloglog 1 30 0.03333333  2.494567e-03 0.14513807
## 5         exact 1 30 0.03333333  8.435709e-04 0.17216946
## 6         logit 1 30 0.03333333  4.675346e-03 0.20200244
## 7        probit 1 30 0.03333333  3.475014e-03 0.16637241
## 8       profile 1 30 0.03333333  3.012987e-03 0.13868254
## 9           lrt 1 30 0.03333333  1.961442e-03 0.13868594
## 10    prop.test 1 30 0.03333333  1.742467e-03 0.19053022
## 11       wilson 1 30 0.03333333  5.908590e-03 0.16670391
```

(とはいえ)DNA鑑定に使ってみよう

- アレル頻度の推定
- 3アレルのマーカー(アレル頻度 (A,B,C)=(0.5,0.3,0.2))
- 6種類のジェノタイプ
- Hardy-Weinberg 平衡

$$\begin{pmatrix} X & A & B & C \\ A & 0.25 & 0.3 & 0.2 \\ B & * & 0.09 & 0.12 \\ C & * & * & 0.04 \end{pmatrix}$$

観測ジェノタイプデータ

$$\begin{pmatrix} X & A & B & C \\ A & 27 & 33 & 20 \\ B & * & 7 & 10 \\ C & * & * & 3 \end{pmatrix}$$

観測ジェノタイプデータ

$$\begin{pmatrix} X & A & B & C \\ A & 27 & 33 & 20 \\ B & * & 7 & 10 \\ C & * & * & 3 \end{pmatrix}$$

A, B, C の観測本数は？

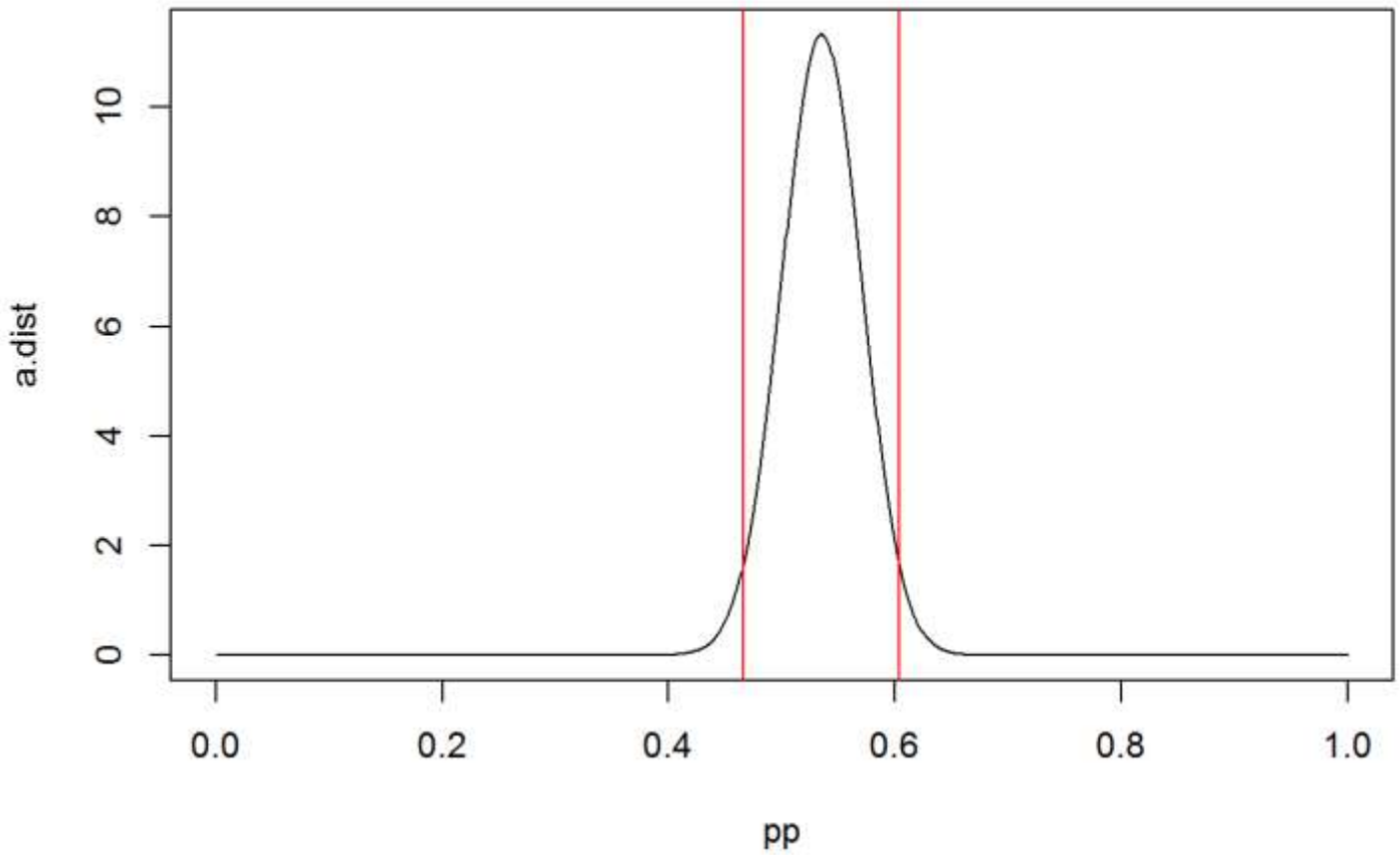
観測ジェノタイプデータ

$$\begin{pmatrix} X & A & B & C \\ A & 27 & 33 & 20 \\ B & * & 7 & 10 \\ C & * & * & 3 \end{pmatrix}$$

A, B, C の観測本数は？

107　57　36

Aアレルの頻度と信頼区間は、A vs non-Aなので、二項分布に基づく方法が使えそう



ディプロタイプ頻度の推定

- AAの人数を元にすれば、
- AA vs. non-AA として、二項分布に基づいて推定できる。

ディプロタイプ頻度の推定

- AAの人数を元にすれば、
- AA vs. non-AA として、二項分布に基づいて推定できる。
- この場合は、HWEを仮定していないことになる。

ディプロタイプ頻度の推定

- AAの人数を元にすれば、
- AA vs. non-AA として、二項分布に基づいて推定できる。
- この場合は、HWEを仮定していないことになる。

- HWEを仮定するべきか、しないべきか、それ「も」問題だ。

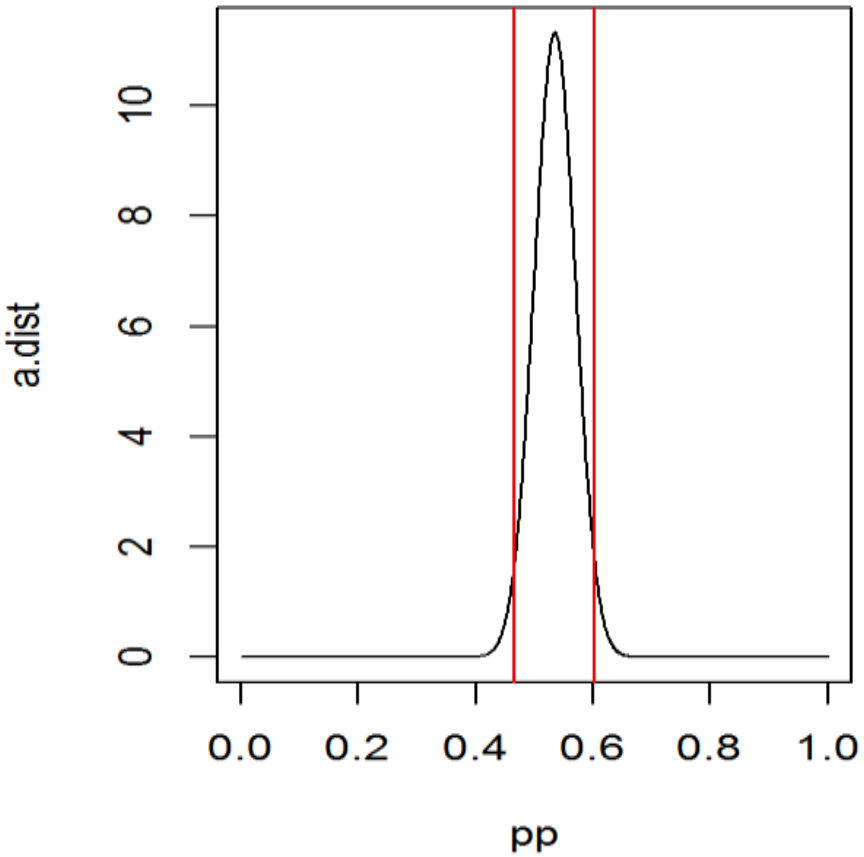
ディプロタイプ頻度の推定

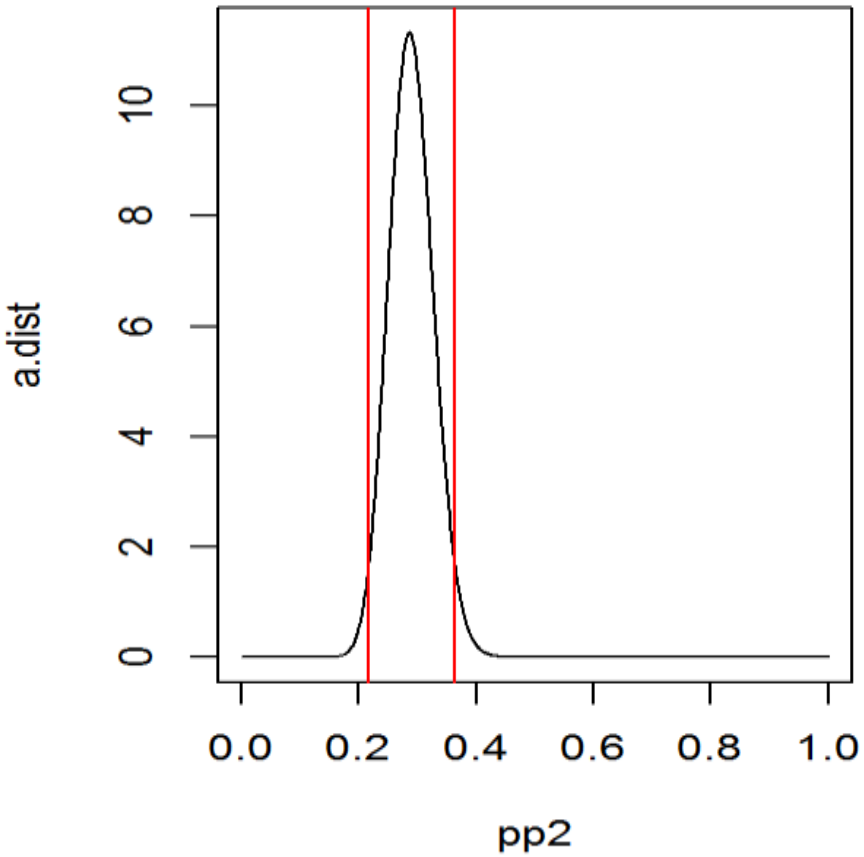
- AAの人数を元にすれば、
- AA vs. non-AA として、二項分布に基づいて推定できる。
- この場合は、HWEを仮定していないことになる。

- HWEを仮定するべきか、しないべきか、それ「も」問題だ。

- が。
- HWEを仮定したとして、アレルAの推定頻度を基に、どうやって、AAディプロタイプの信頼区間推定をするのか？
- AAの頻度はアレル頻度の２乗なので…

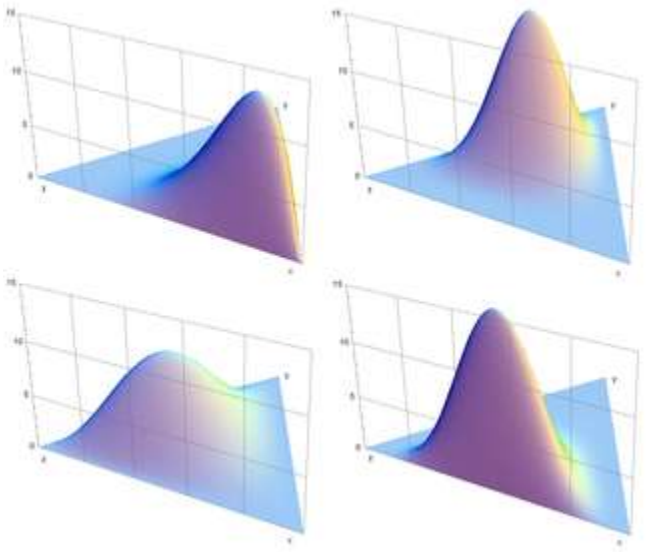
横軸を p -> p x p に変換する？

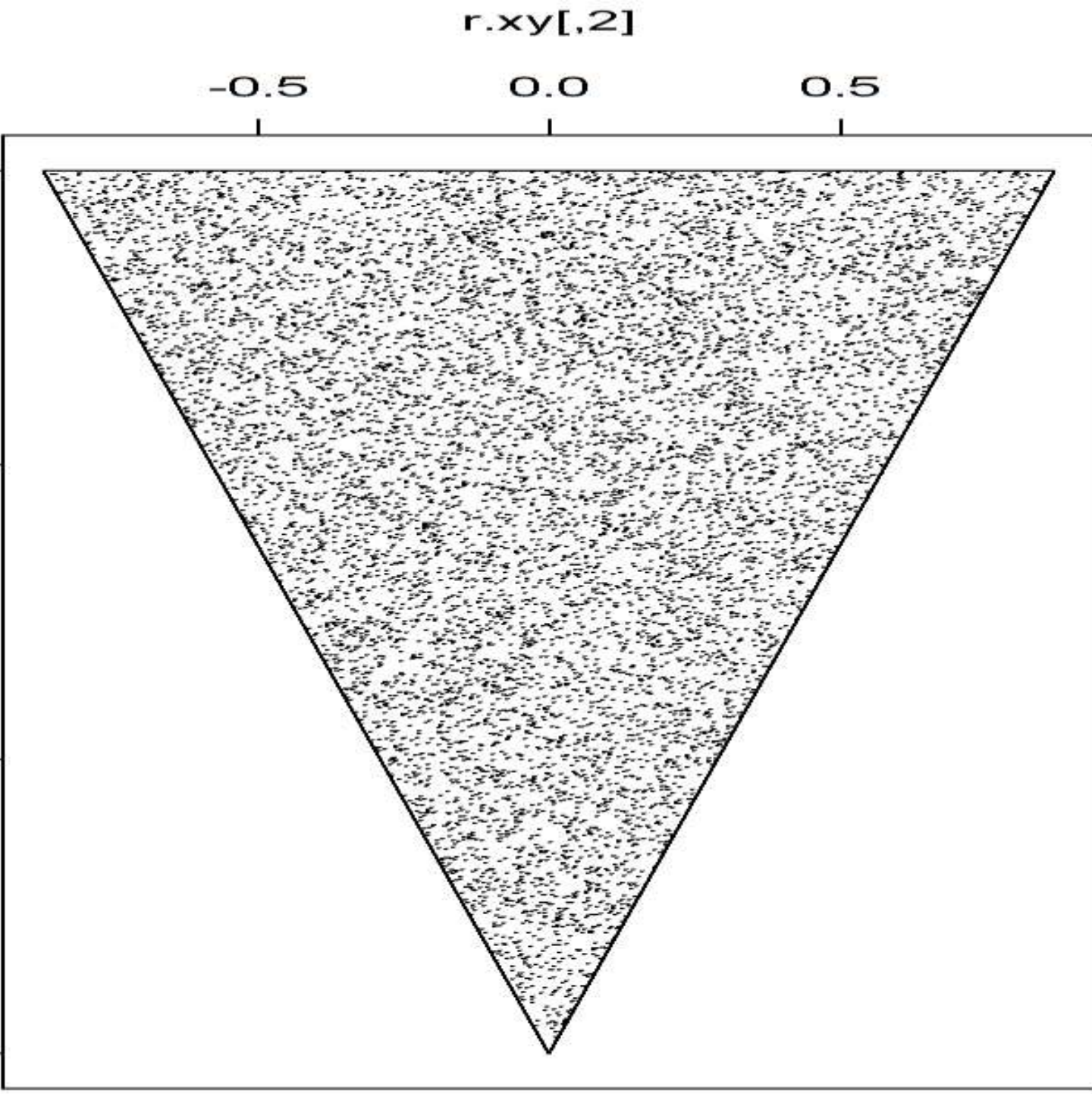


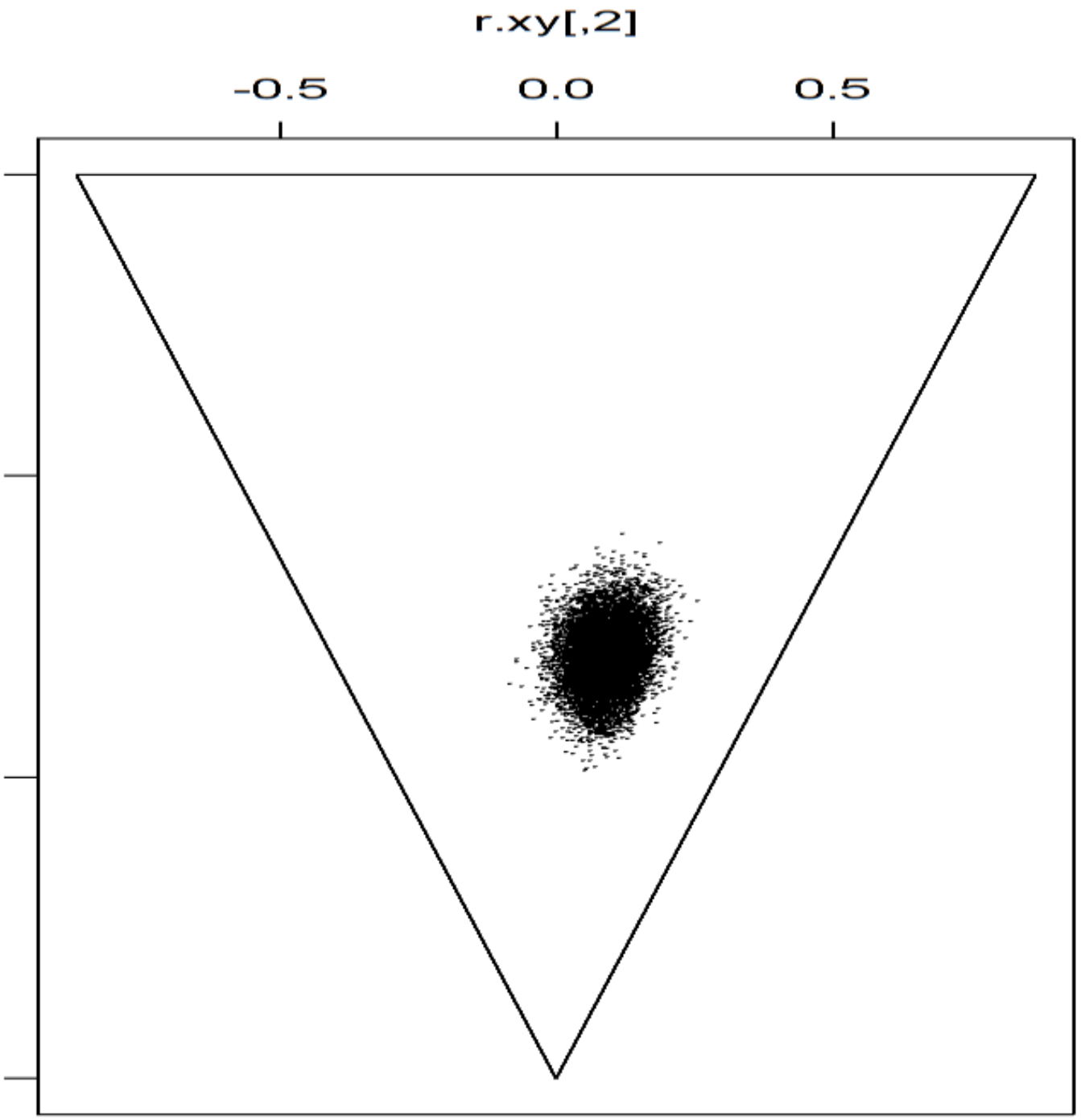


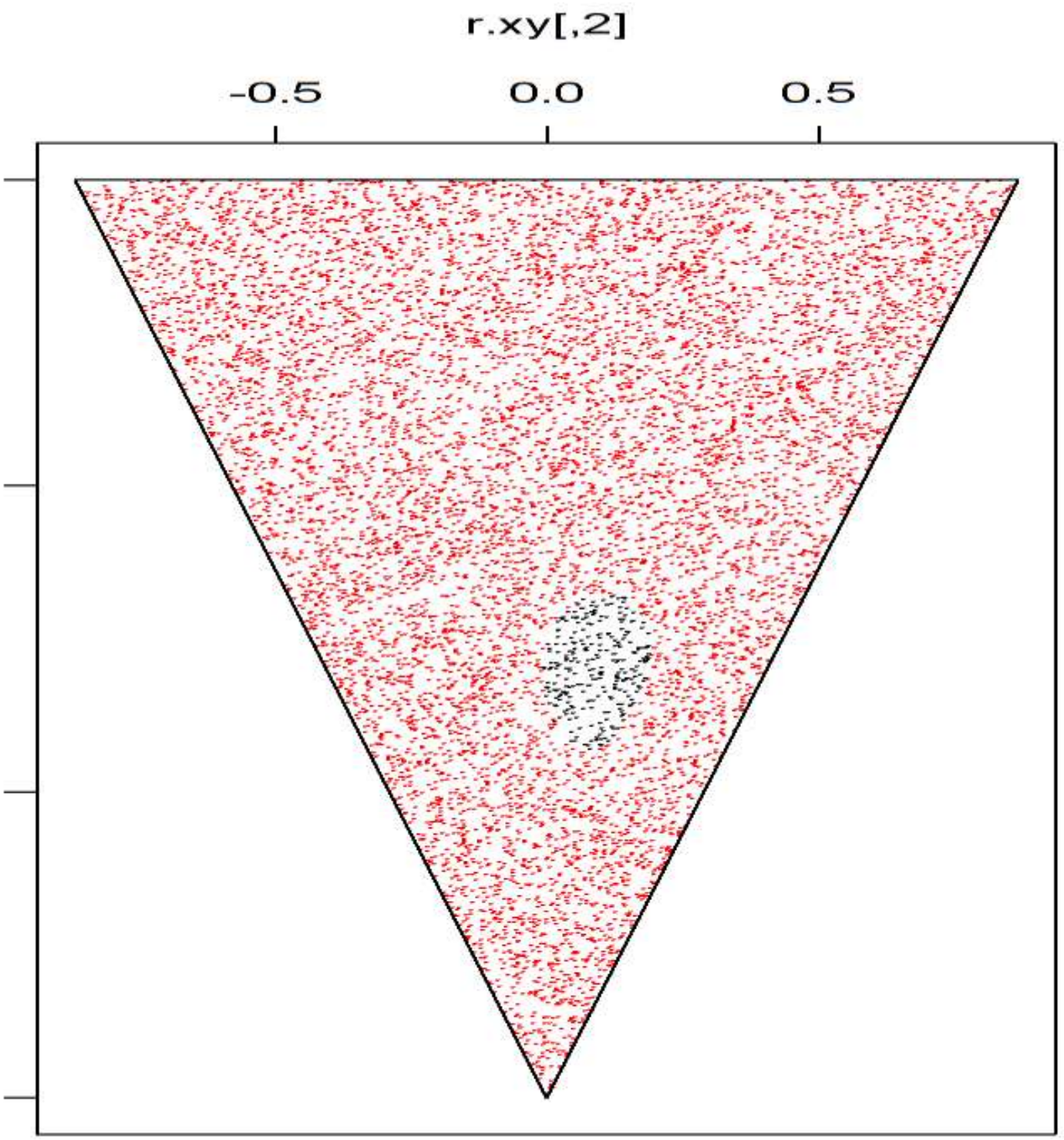
ABの頻度はどうする？

- アレルAの頻度とアレルBの頻度をそれぞれ求める？
- アレルAの頻度が高いとき、アレルBの頻度は低いはず。
- お互いに影響し合っているので、別々に推定したり、別々の信頼区間を考えるのはまずい。
- 多項分布のベイズ推定はディリクレ分布
- A + B + C = 1 を満足する自由度2の分布




r.xy[,2]
-0.5
0.0
0.5


r.xy[,2]
-0.5
0.0
0.5


r.xy[,2]
-0.5
0.0
0.5

乱数を使ってみては？

- 今回のように、ベータ分布・ディリクレ分布などを使うこともできる
- もっと、複雑になってくると、「〇〇分布」は使えない
- そんなときは乱数をつかったり、リサンプリングを行ったりする

DNA鑑定での尤度比

- 容疑者のジェノタイプが現場の試料のそれと一致したとき。
- たまたま、一致したのか、同一人物だから一致したのかは、それぞれの仮説の尤度の比で計算する。
- 同一人物の場合の尤度は1だから、たまたまの場合の尤度を計算すればよい。

DNA鑑定での尤度比

- マーカーごとの観察が独立とみなせるならば、個々のマーカーでの尤度の積。
- 複数のマーカー、それぞれのマーカーのアレル数を適当に与えてシミュレーションしてみる。
- ディプロタイプのデータベースをシミュレーション作成する。

```
## 2.5% 97.5% ## 5.340572e-07 1.358108e-06
  ##
```

たとえば

- マーカー数５
- アレル数　3〜6
- 各マーカーのジェノタイプが、最頻アレルと第二最頻アレルのホモ接合型であるような場合
- 95% 区間推定値
  - 5.340572e-07 1.358108e-06

今日、触れなかったこと

- ２つの仮説から尤度が出て、その比を問題にするとき
  - 片方の尤度が高いときに
    - もう片方の尤度が高い場合と低い場合とを考慮
  - その逆も
- ある仮説が真であるとみなしたときに、別の仮説は真ではなくなる。その相互作用を考えるとき。しかもそれが多人数に及ぶとき

- そもそも「事前分布」をどうするのがよいのかは、統計学的に未解決の問題
  - 「成功率」の事前分布は、一様分布ではない(かもしれない)…

本日のスライド、資料

- http://statgenet-kyotouniv.wikidot.com/handouts-slides
  - 尤度比の信頼区間(法数学勉強会2016年6月)