ディリクレ分布
と
ディリクレ過程

法数学勉強会

2017/10/14

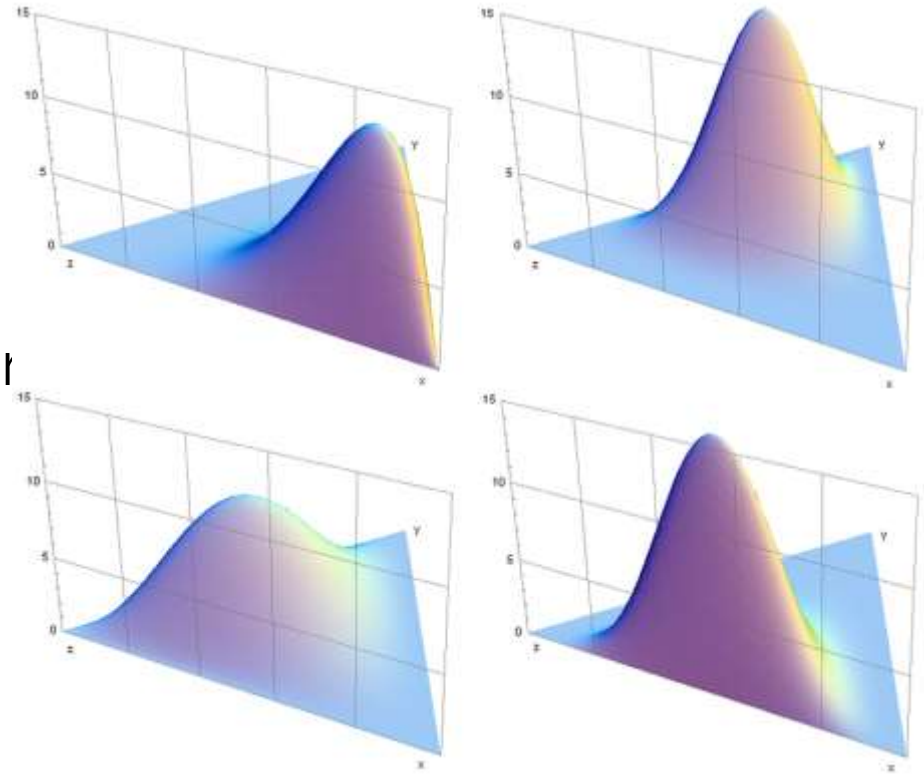
京都大学(医)統計遺伝学分野

山田　亮

今日と次回(と次々回)の構成

- ディリクレ分布

- ディリクレ過程
  - https://en.wikipedia.org/wiki/Dirichlet_pr
    urant_process
- (ノンパラメトリック・ベイズ)



どうしてディリクレ分布

どうしてディリクレ分布

- アレル頻度のためにある

どうしてディリクレ過程

どうしてディリクレ過程

- アレルの種類数が未知なままに、アレル頻度の推定をしたいことがある
  - Y染色体ハプロタイプ頻度推定

ヨハン・ペーター・グスタフ・ルジューヌ・ディリクレ

- https://ja.wikipedia.org/wiki/%E3%83%9A%E3%83%BC%E3%82%BF%E3%83%BC%E3%83%BB%E3%82%B0%E3%82%B9%E3%82%BF%E3%83%95%E3%83%BB%E3%83%87%E3%82%A3%E3%83%AA%E3%82%AF%E3%83%AC

どうしてディリクレ分布

- アレル頻度のためにある
- １０人のSNP(一塩基多型）のジェノタイプを調べたら
  - AA 5人
  - AG 3人
  - GG 2人
- だった

どうしてディリクレ分布

- アレル頻度のためにある
- １０人のSNP(一塩基多型）のジェノタイプを調べたら
  - AA 5人
  - AG 3人
  - GG 2人
- だった
- アレルAの頻度はいくつ？

標本頻度と母集団頻度

- AA 5人
- AG 3人
- GG 2人

- A : 2 * 5 + 1 * 3 + 0 * 2 = 13
- G : 0 * 5 + 1 * 3 + 2 * 2 = 7

- 13 + 7 = 20 = 10 * 2

標本頻度と母集団頻度

- AA 5人
- AG 3人
- GG 2人

- A : 2 * 5 + 1 * 3 + 0 * 2 = 13
- G : 0 * 5 + 1 * 3 + 2 * 2 = 7

- 13 + 7 = 20 = 10 * 2

- Aの割合は
  - 13/20 = 0.65

標本頻度

- Aの割合は
  - 13/20 = 0.65
- 0.65 の意味することは
  - 20本の染色体を観測したら、その65%がAだった、という事実

母集団の頻度

- ある人 X のあるマーカーのジェノタイプを調べたら
  - AAだった
- 現場の試料 Y のジェノタイプを調べたら
  - AAだった
- Xが現場に試料を残したと仮定したとき
  - 尤度 $L(Y=AA \mid X=AA) = 1$; $Y = X$

母集団の頻度

- ある人 X のあるマーカーのジェノタイプを調べたら
  - AAだった
- 現場の試料 Y のジェノタイプを調べたら
  - AAだった
- Xではない、誰か Z が現場に試料を残したと仮定したとき
  - 尤度 L(Y=AA | Z=??) = ? ; Y = Z

母集団の頻度

- ある人 X のあるマーカーのジェノタイプを調べたら
  - AAだった
- 現場の試料 Y のジェノタイプを調べたら
  - AAだった
- Xではない、誰か Z が現場に試料を残したと仮定したとき
  - 尤度 L(Y=AA | Z=??) = ? ; Y = Z
  - Pr(Z=AA), Pr(Z=AG), Pr(Z=GG) が知りたい
  - なぜなら
    - L(Y=AA |Z=??) = Pr(Z=AA) * L(Y=AA|Z=AA) + Pr(Z=AG) * L(Y=AA|Z=AG) + Pr(Z=GG) * L(Y=AA| Z=GG)
    - L(Y=AA | Z=??) = Pr(Z=AA) * 1 + 0 + 0 = Pr(Z=AA)

母集団の頻度

- L(Y=AA | Z=??) = Pr(Z=AA)
- Pr(Z=AA), Pr(Z=AG), Pr(Z=GG) が知りたい
- Pr(Z=**)は、想定している集団のジェノタイプ頻度

- Pr(Z=AA) = ???? 0.65 * 0.65 ????

- AA 5人
- AG 3人
- GG 2人

- A : 2 * 5 + 1 * 3 + 0 * 2 = 13
- G : 0 * 5 + 1 * 3 + 2 * 2 = 7

- 13 + 7 = 20 = 10 * 2

- Aの割合は
  - 13/20 = 0.65

母集団の頻度

- AA 5人
- AG 3人
- GG 2人

- A : 2 * 5 + 1 * 3 + 0 * 2 = 13
- G : 0 * 5 + 1 * 3 + 2 * 2 = 7

- 13 + 7 = 20 = 10 * 2

- Aの割合は
  - 13/20 = 0.65

- AA 5000人
- AG 3000人
- GG 2000人

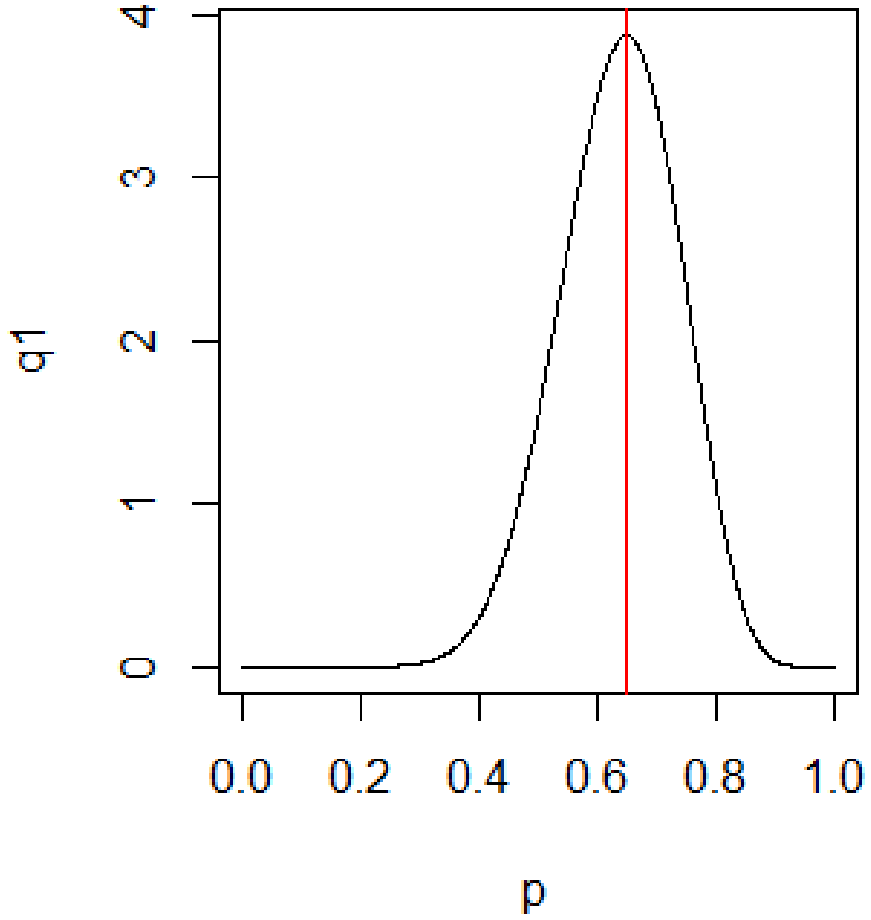
- A : 2 * 5000 + 1 * 3000 + 0 * 2000 = 13000
- G : 0 * 5000 + 1 * 3000 + 2 * 2000 = 7000

- 13000 + 7000 = 20000 = 10000 * 2

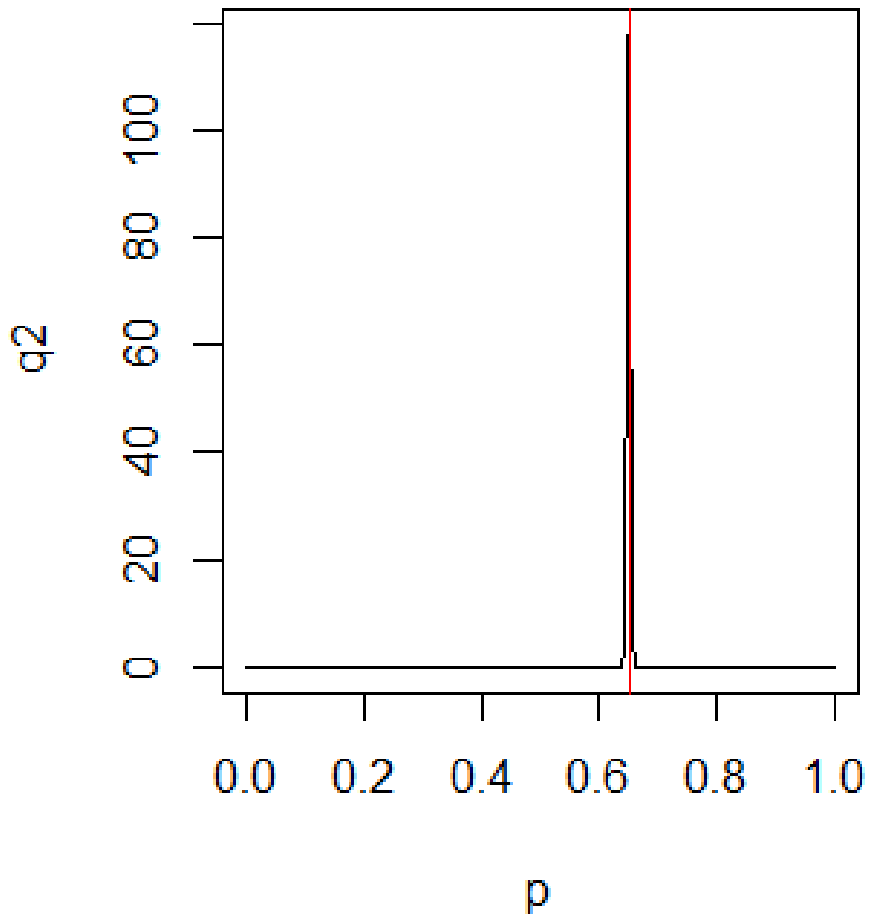
- Aの割合は
  - 13000/20000 = 0.65

母集団の頻度のベイズ推定

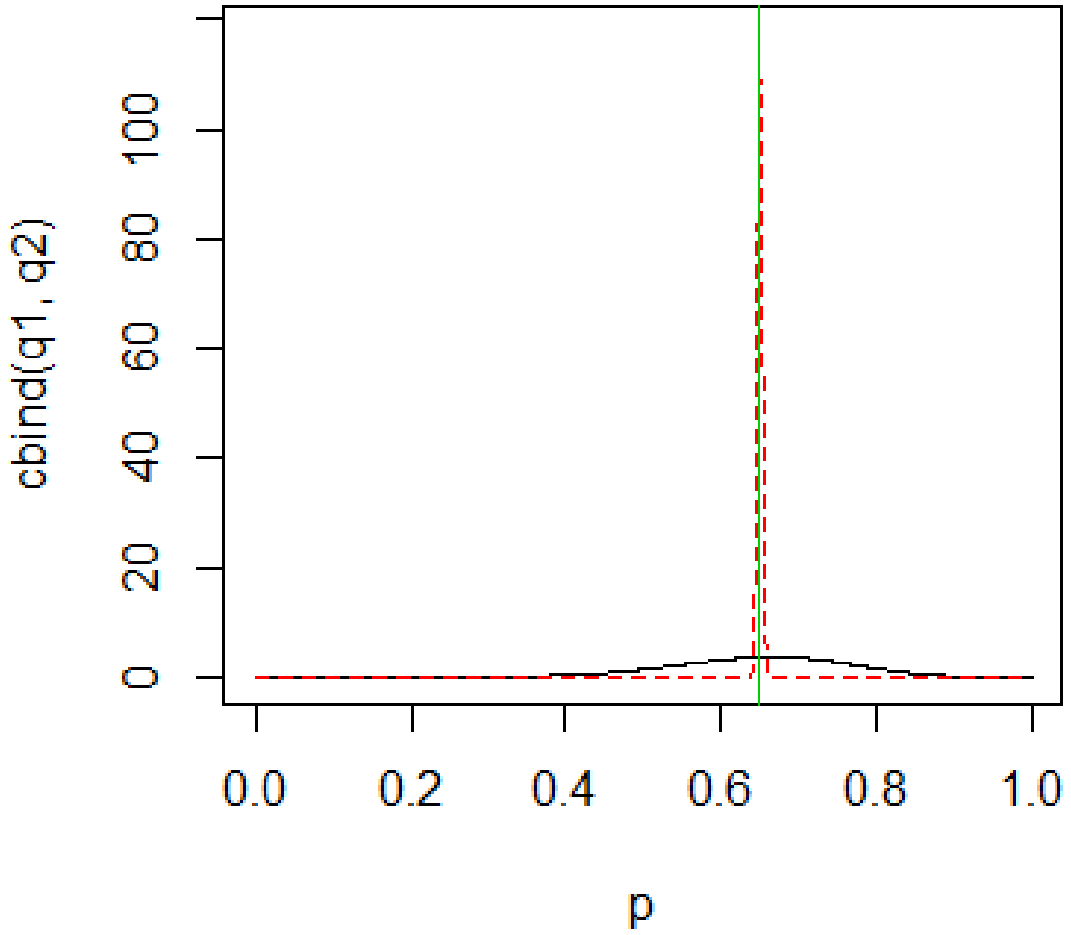
全10人



全10000人



母集団の頻度のベイズ推定



尤度を計算する

- 母集団でのAの割合が p だとしたとき
- 20本中13本がAとして観察される確率は
- 20本中13本がAという観測をしたとき
- 母集団でのAの割合が p である尤度

- $\frac{20!}{13!7!} \times p^{13} \times (1-p)^{7}$

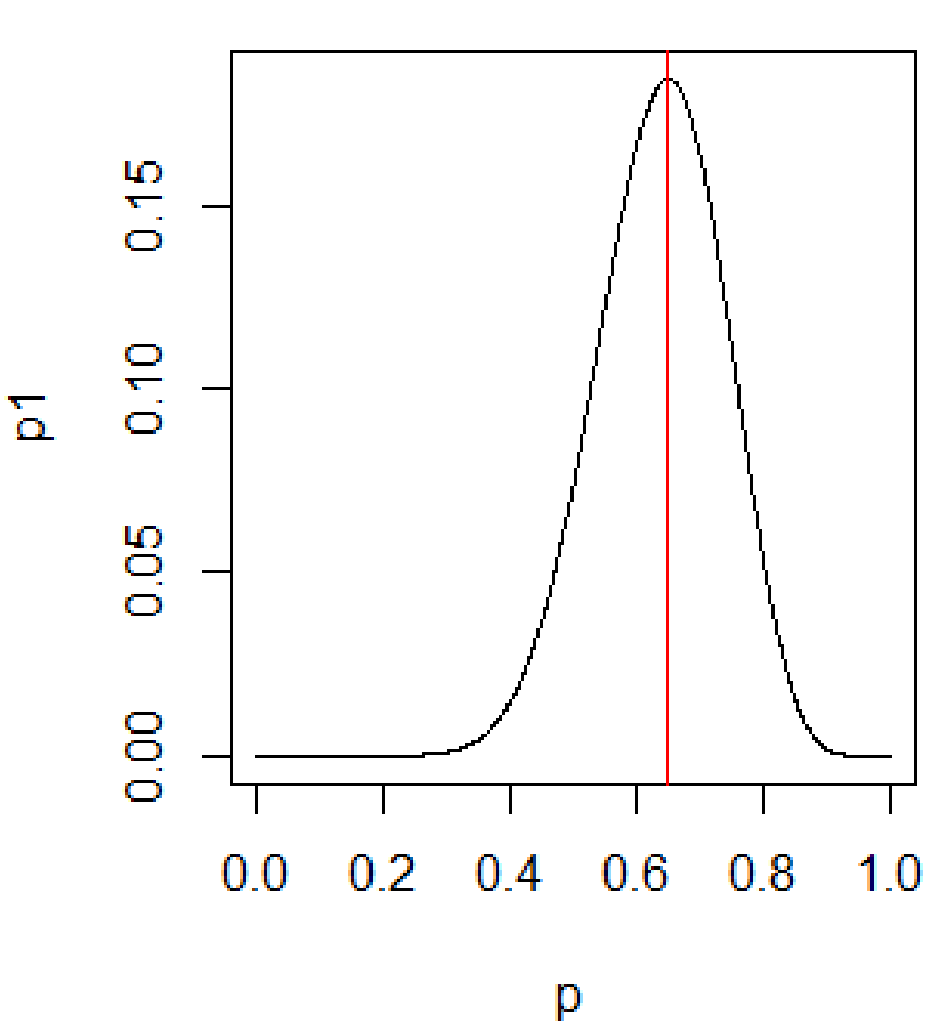
- AA 5人
- AG 3人
- GG 2人

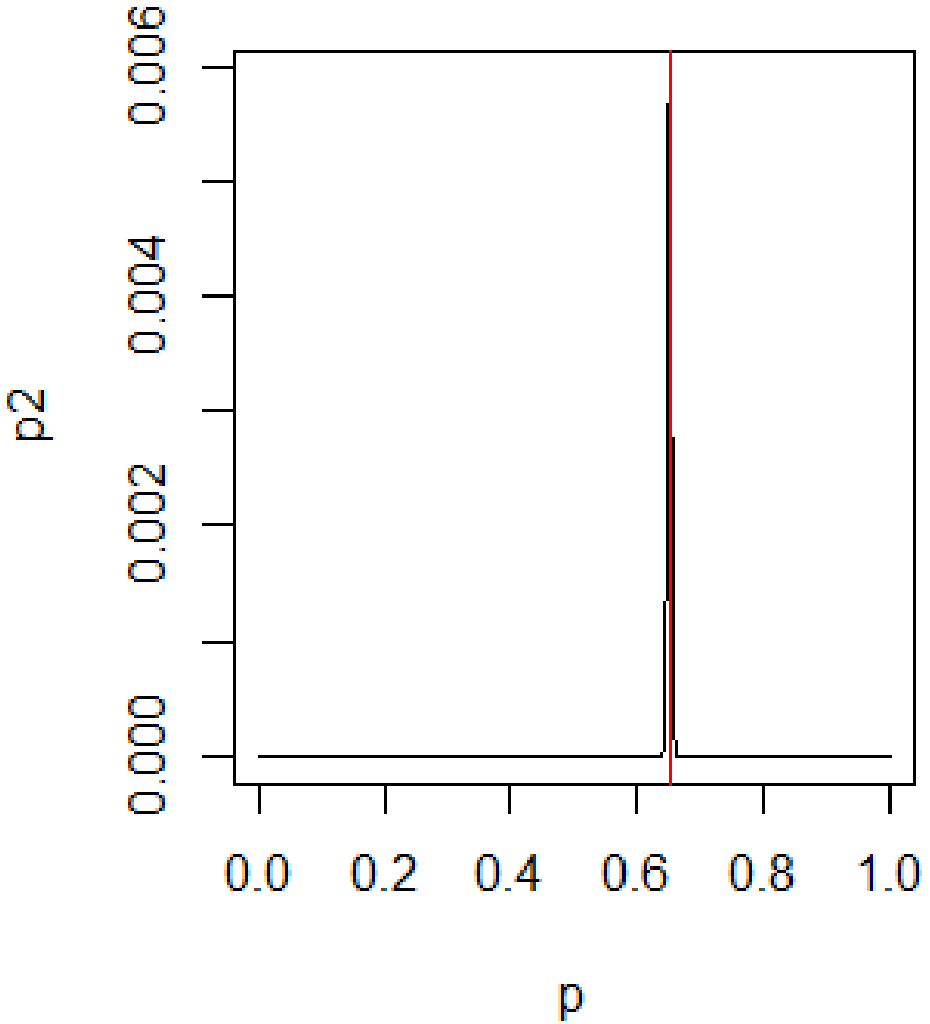
- A : 2 * 5 + 1 * 3 + 0 * 2 = 13
- G : 0 * 5 + 1 * 3 + 2 * 2 = 7

- 13 + 7 = 20 = 10 * 2

- Aの割合は
  - 13/20 = 0.65

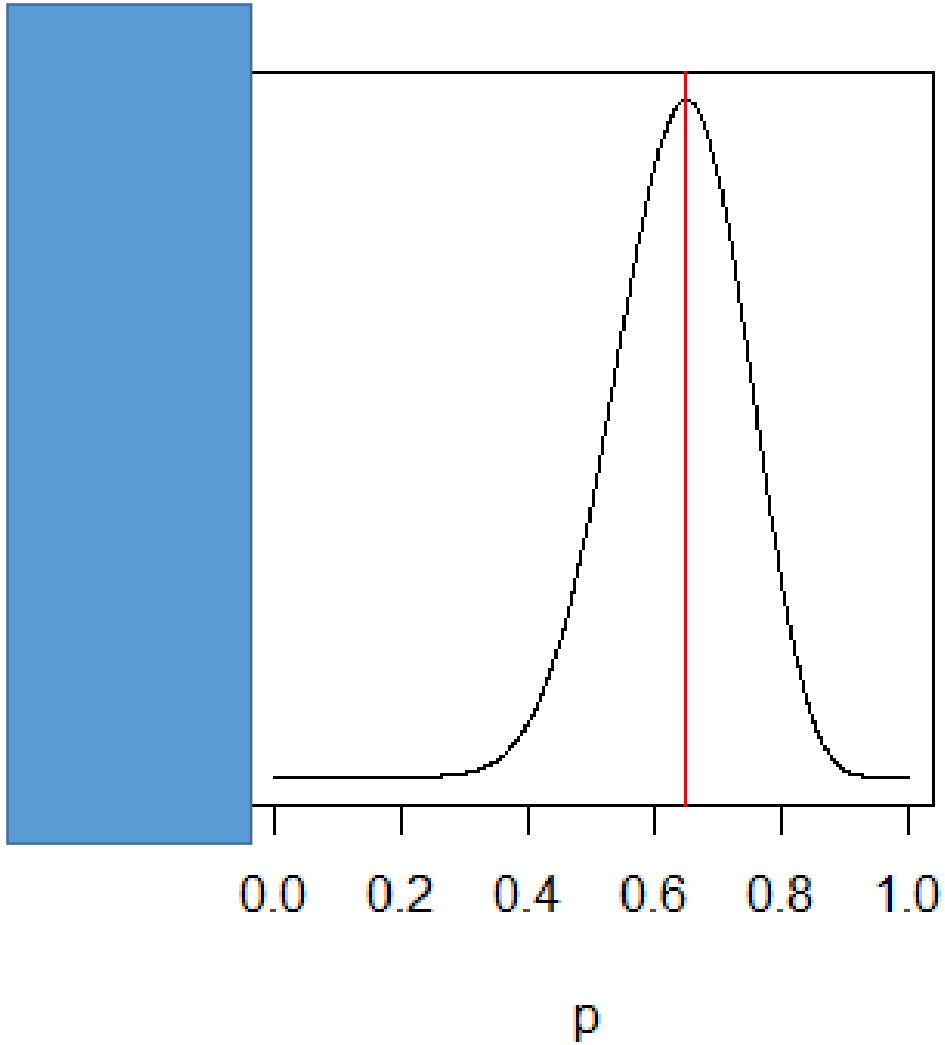
- $\frac{20!}{13!7!} \times p^{13} \times (1-p)^{7}$
- $\frac{20000!}{13000!7000!} \times p^{13000} \times (1-p)^{7000}$

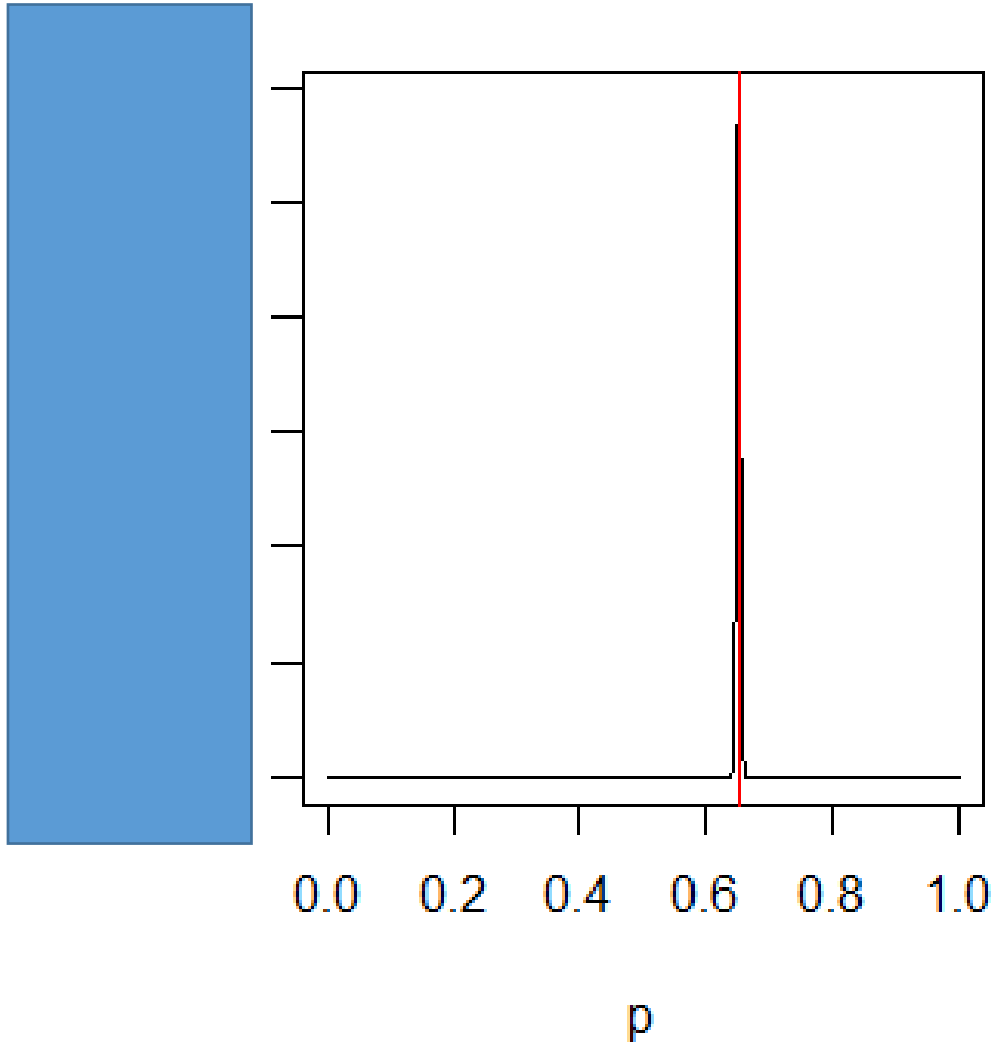




- $\frac{20!}{13!7!} \times p^{13} \times (1-p)^{7}$
- $\frac{20000!}{13000!7000!} \times p^{13000} \times (1-p)^{7000}$

尤度関数は積分して１になるとは限らない





尤度関数を定数倍して、積分＝１にする

- $? \times\ p^{13} \times (1-p)^{7}$
- p について0から1まで積分して、1になるように？を掛ける

- $\frac{20!}{13!7!} \times\ p^{13} \times (1-p)^{7}$　確率・尤度

- $\frac{21!}{13!7!} \times\ p^{13} \times (1-p)^{7}$　積分して１になるように直した

どこが違う？

ベータ分布

- 起きるか起きないか、の観測データに基づいて、成功率の尤度を考える
- この尤度関数を定数倍して、積分が 1 になるような確率密度分布をベータ分布と呼ぶ
- 「起きた回数：S」「起きなかった回数：F」
- $\frac{(S+F+1)!}{S!F!} \times p^S \times (1-p)^F$

ベータ分布

- $\frac{(S+F+1)!}{S!F!} \times p^S \times (1-p)^F$
- 別の書き方もよくする
- S -> A-1, F -> B-1
- $\frac{((A-1)+(B-1)+1)!}{(A-1)!(B-1)!} \times p^{(A-1)} \times (1-p)^{(B-1)}$
- $\Gamma(z+1) = z!$, $\Gamma(w) = (w-1)!$
- $\frac{\Gamma(A+B)!}{\Gamma(A)!\,\Gamma(B)!} \times p^{(A-1)} \times (1-p)^{(B-1)}$

練習問題

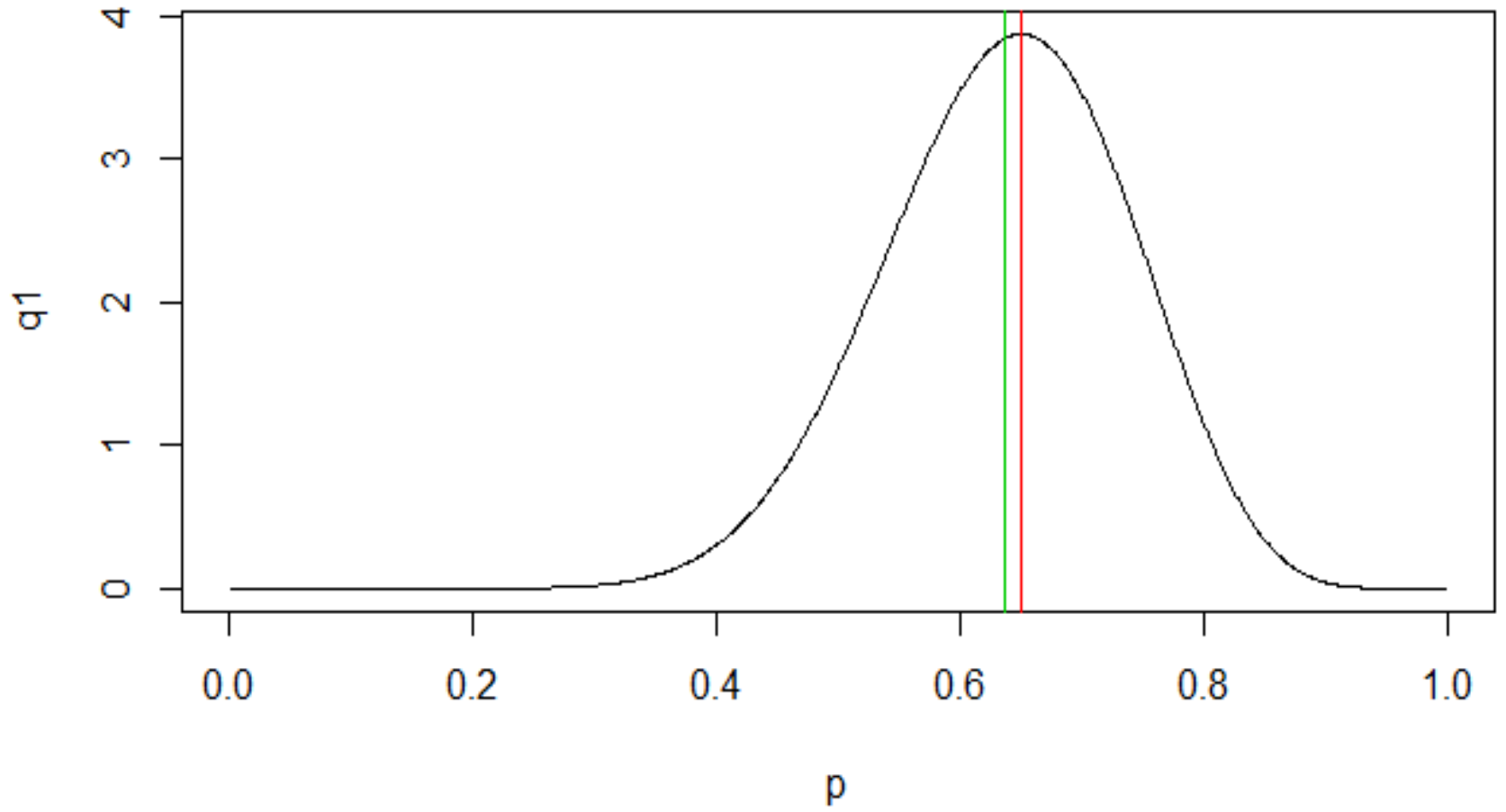
- 成功率が a であるような確率事象を 5 回の試みたところ、４回成功して1回失敗する確率を計算する式を書け

- 5回の試行で4回成功し、1回失敗した。この確率事象の成功率が a であるような尤度を計算する式を書け

- このような尤度に比例したベータ分布の確率密度関数の式を書け

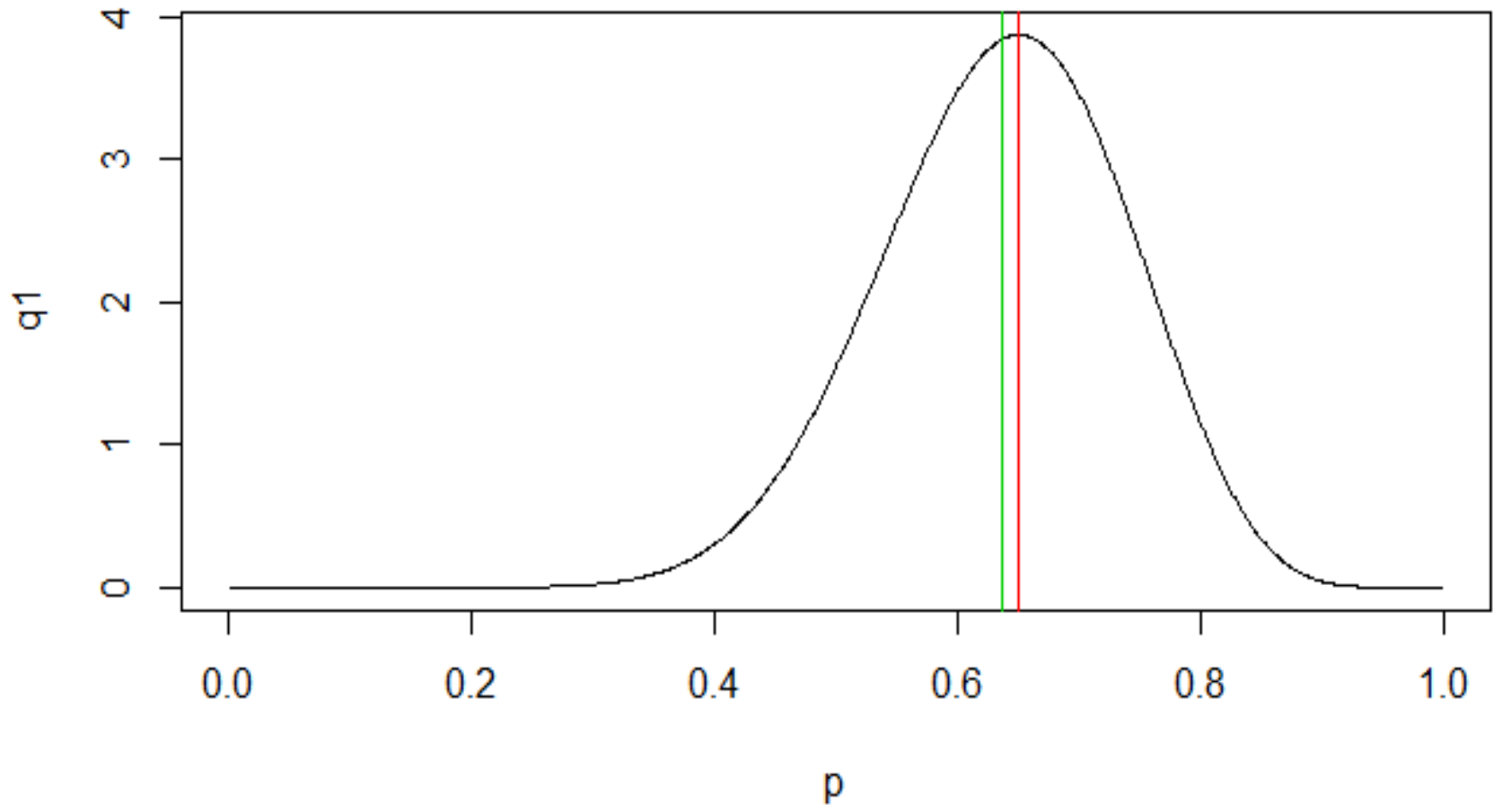
ベータ分布の期待値と最頻値

- 赤：最頻値
- 緑：期待値



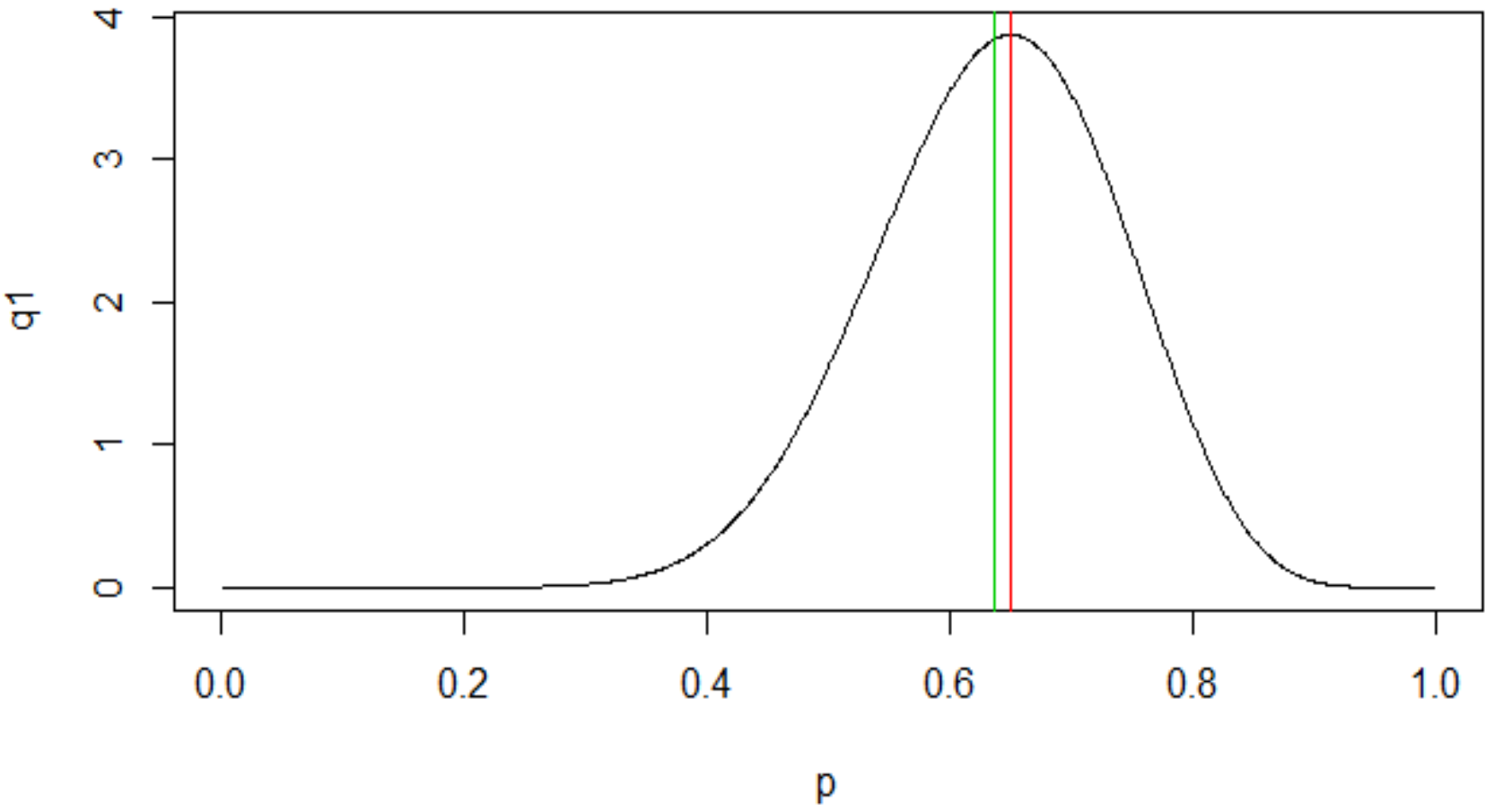
ベータ分布の期待値と最頻値

- 赤：最頻値
- 尤度が最大
- 最尤推定値
- 13/20
- 標本頻度



ベータ分布の期待値と最頻値

- 緑：期待値
- 最頻値よりも
- 中央寄り
- 20回の試行後
- 21回目に成功する確率



- $\beta'(\mathrm{p}|\mathrm{S},\mathrm{F}) = \frac{(S+F+1)!}{S!F!} \times p^S \times (1-p)^F$

練習問題

- $\beta\,(p|A,B) = \frac{\Gamma(A+B)!}{\Gamma(A)!\,\Gamma(B)!} \times p^{(A-1)} \times (1-p)^{(B-1)}$

- 10回の試行、４回の成功
  - 式でのS,Fはいくつか、A,Bはいくつか？
  - 「標本成功率」はいくつか？
  - この確率事象の成功率はベータ分布を取る。そのベータ分布の確率密度を最大とする成功率はいくつか？
  - この確率事象の成功率の最尤推定値はいくつか？
  - この確率事象の成功率を表すベータ分布の期待値はいくつか？

- $\beta'(\mathrm{p}|\mathrm{S},\mathrm{F}) = \frac{(S+F+1)!}{S!F!} \times p^S \times (1-p)^F$

練習問題

- $\beta\,(p|A,B) = \frac{\Gamma(A+B)!}{\Gamma(A)!\,\Gamma(B)!} \times p^{(A-1)} \times (1-p)^{(B-1)}$

- 0回の試行、0回の成功
  - 式でのS,Fはいくつか、A,Bはいくつか？
  - 「標本成功率」はいくつか？
  - この確率事象の成功率はベータ分布を取る。そのベータ分布の確率密度を最大とする成功率はいくつか？
  - この確率事象の成功率の最尤推定値はいくつか？
  - この確率事象の成功率を表すベータ分布の期待値はいくつか？
- 一様分布
  - 一様分布となるS,Fはいくつか、A,Bはいくつか？

事前分布

事後分布

共役事前分布

- $\beta'(\mathrm{p}|\mathrm{S},\mathrm{F}) = \frac{(S+F+1)!}{S!F!} \times p^S \times (1-p)^F$
- $\beta\,(p|A,B) = \frac{\Gamma(A+B)!}{\Gamma(A)!\,\Gamma(B)!} \times p^{(A-1)} \times (1-p)^{(B-1)}$

- 実験前に、成功率 p が一様分布であると思っていた
  - p = 0.3の事前確率はいくつか？p = 0.4の事前確率いくつか？
  - 事前分布をβ(p|A,B)と考えると、A,Bをいくつとみなしていたことになるか？
- 実験を行った。試行3回、成功2回だった
  - p = 0.3とp = 0.4との尤度比をβ分布関数を用いて考えよ
  - 事後確率＝事前確率 x 尤度である。どのような式になるか？
- 実験前に、なんとなく、成功率が0.5より高いような予感がしていた、と言う
  - その事前分布をpの関数として図示してみよ
  - その事前分布に近いベータ分布のパラメタがa,bだったとする
  - このとき、事後分布はどのような式になるか？

共役事前分布

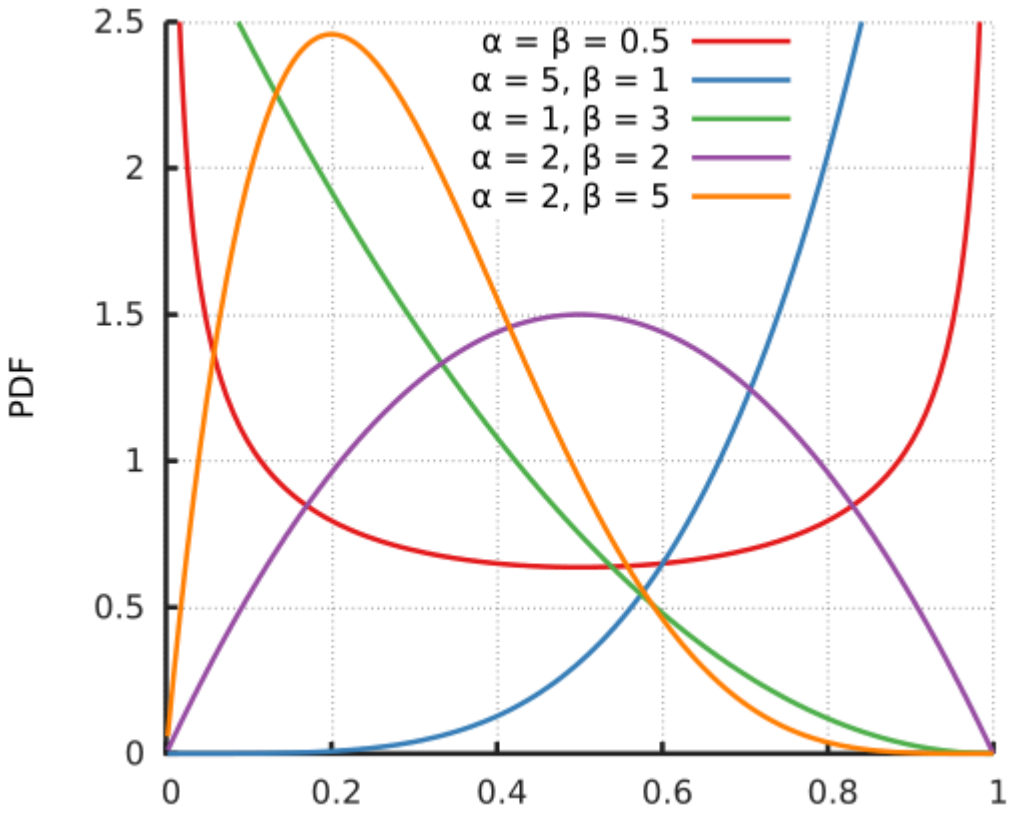
- 成功確率をベイズ推定する場合
- 事前分布としてベータ分布を用いると
- 実験後の事後分布は
  - 事前分布のベータ分布を表すパラメタ値と
  - 尤度を表すベータ分布を表すパラメタ値と
  - だけを使って
  - 事後分布のベータ分布を表すパラメタ値が算出できる
- このような便利な事前分布を共役事前分布と呼ぶ
  - したがって
  - ベータ分布は、二項現象にとっての共役事前分布
  - ベータ分布と二項分布は共役

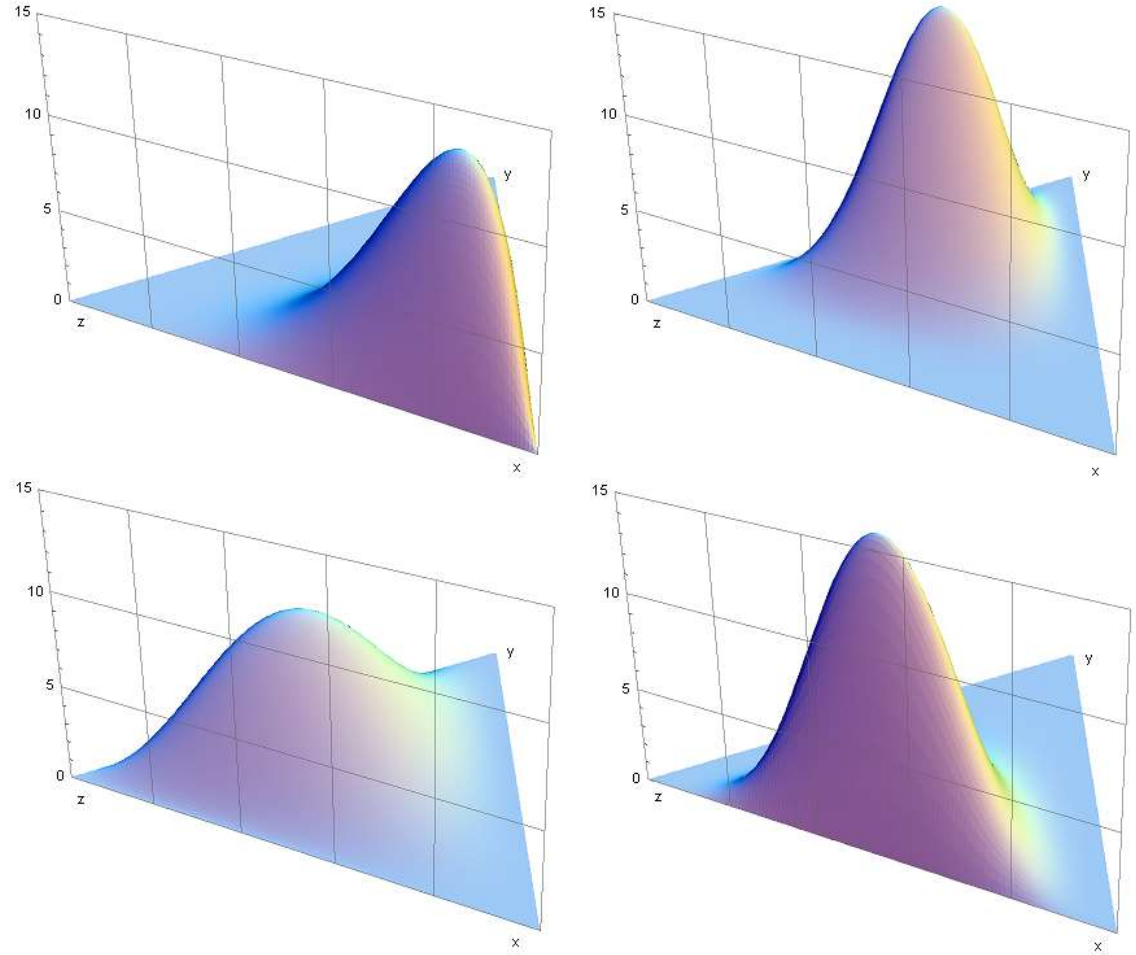
ディリクレ分布は、ベータ分布の多項版

- ベータ分布は、二項事象の成功確率の尤度関数に比例し、積分して１になるような確率密度分布

- ディリクレ分布は、項数が増えたときに、同じく
  - 尤度関数に比例し
  - 積分して１になるような確率密度関数

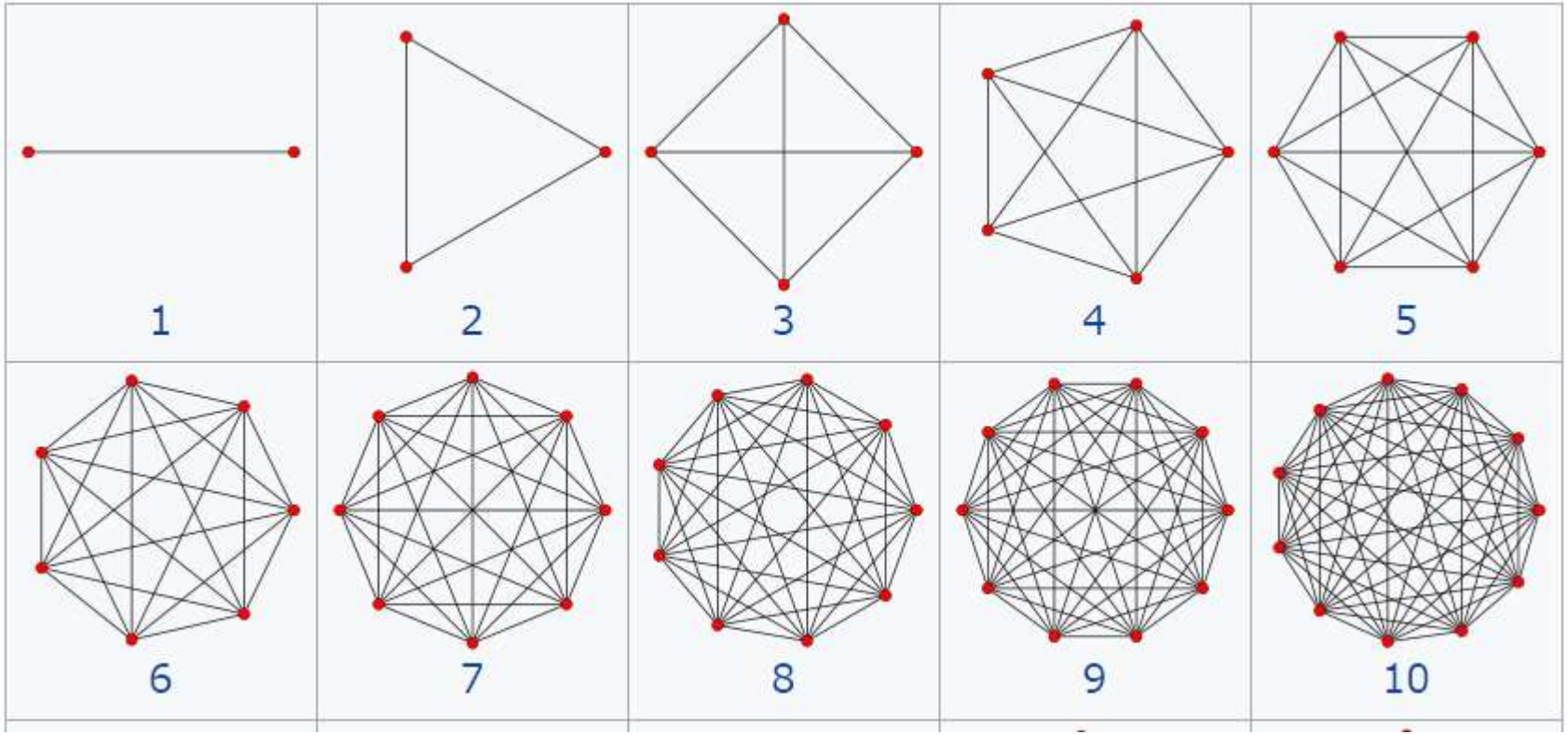
ベータ分布とディリクレ分布(項数３)の形



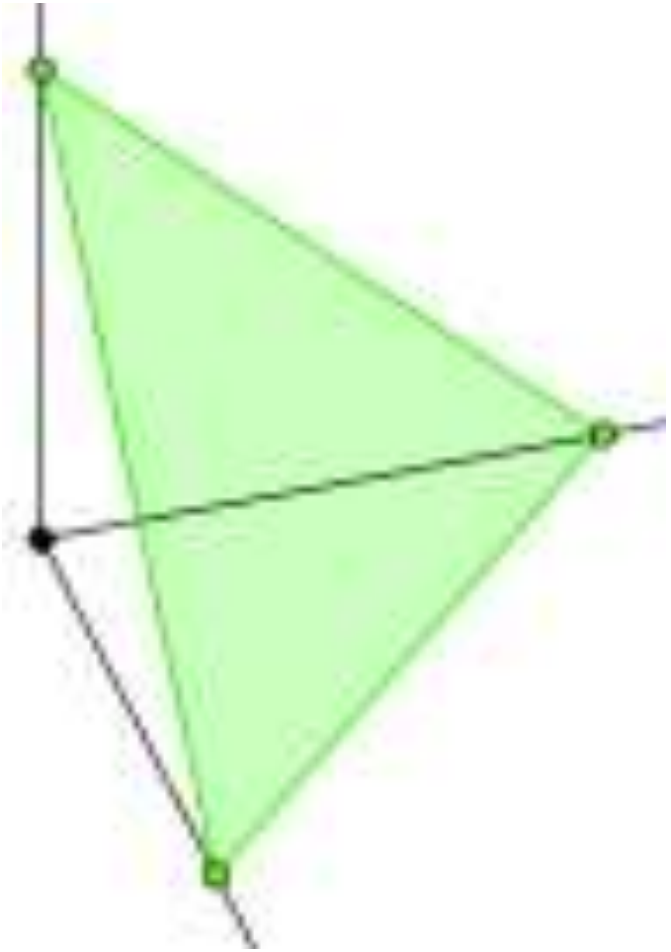


正単体
ベータ分布・ディリクレ分布の「領域」

- 三角形の一般化



- p1 + p2 + … + pk = 1の一般化



式を比べる
K=2がベータ分布になることを確認

- ベータ分布 $\beta(p|A,B) = \frac{\Gamma(A+B)!}{\Gamma(A)!\,\Gamma(B)!} \times p^{(A-1)} \times (1-p)^{(B-1)}$

台	$x \in [0;1]$

- ディリクレ分布

台	$x_1, \ldots, x_K$ 、ここで、 $x_i \in [0,1]$ かつ $\sum x_i = 1$

式を比べる

$$\beta\,(p|A,B) = \frac{\Gamma(A+B)!}{\Gamma(A)!\,\Gamma(B)!} \times p^{(A-1)} \times (1-p)^{(B-1)}$$

- ベータ分布

確率密度関数

$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

- ディリクレ分布

確率密度関数

$$\frac{1}{\mathrm{B}(\boldsymbol{\alpha})}\prod_{i=1}^{K} x_i^{\alpha_i-1}$$

ここで、 $\mathrm{B}(\boldsymbol{\alpha}) = \dfrac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}$

$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$

式を比べる

$$\beta\,(p|A,B) = \frac{\Gamma(A+B)!}{\Gamma(A)!\,\Gamma(B)!} \times p^{(A-1)} \times (1-p)^{(B-1)}$$

- ベータ分布

期待値

$$\mathrm{E}[X] = \frac{\alpha}{\alpha+\beta}$$

- ディリクレ分布

期待値

$$\mathrm{E}[X_i] = \frac{\alpha_i}{\sum_k \alpha_k}$$

式を比べる

$$\beta\,(p|A,B) = \frac{\Gamma(A+B)!}{\Gamma(A)!\,\Gamma(B)!} \times p^{(A-1)} \times (1-p)^{(B-1)}$$

- ベータ分布

最頻値	$\dfrac{\alpha-1}{\alpha+\beta-2}$

- ディリクレ分布

最頻値	$x_i = \dfrac{\alpha_i - 1}{\sum_{i=1}^{K} \alpha_i - K}$

式を比べる

$$\beta\,(p|A,B) = \frac{\Gamma(A+B)!}{\Gamma(A)!\,\Gamma(B)!} \times p^{(A-1)} \times (1-p)^{(B-1)}$$

- ベータ分布
  - 一様分布のα。Bは？

- ディリクレ分布
  - 一様分布のα＝(α1,α2,…)は？

式を比べる

最頻値	$x_i = \dfrac{\alpha_i - 1}{\sum_{i=1}^{K} \alpha_i - K}$

最頻値	$\dfrac{\alpha - 1}{\alpha + \beta - 2}$

ディリクレ過程

- 次回のお楽しみ

中華料理店過程(チャイニーズレストラン過程)