行動科学演習・数理行動科学研究演習 『社会科学のためのベイズ統計モデリング』第 8 章

Hiroshi Hamada
Tohoku University

Jul, 2021
at Tohoku University

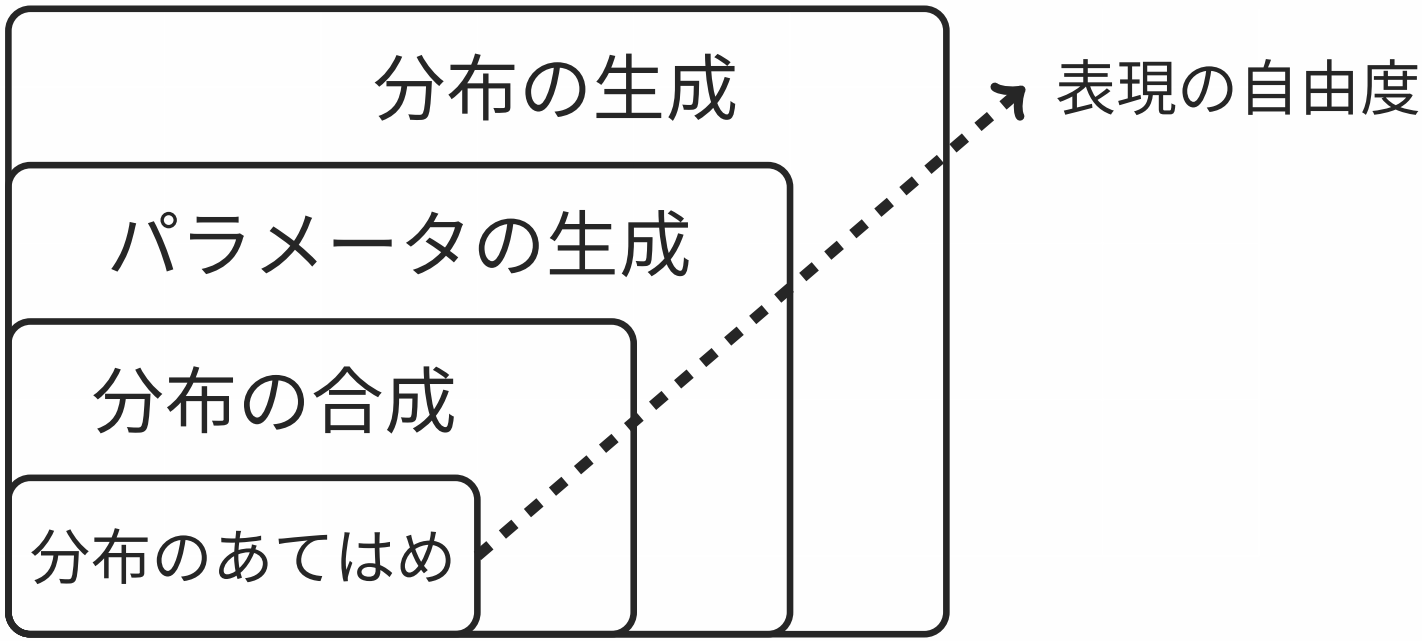

Figure: 確率モデルの作り方のタイプ

分布をあてはめるモデル

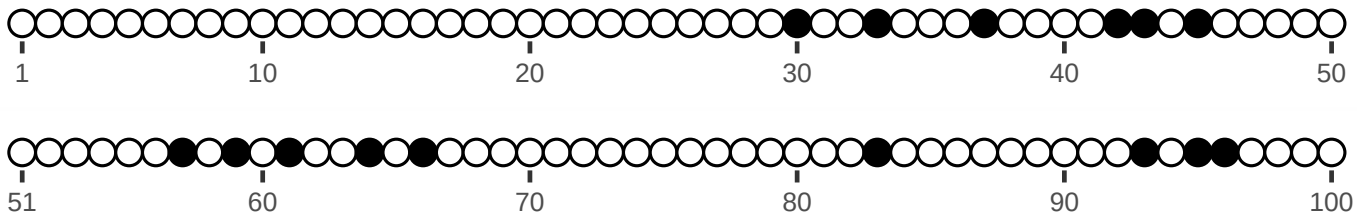

Figure: 藤井聡太七段の 100 局経過時点での通算勝敗記録

モデル 1

$$Y_i \sim \mathrm{Bernoulli}(q), \quad i = 1, \ldots, n$$
$$q \sim \mathrm{Beta}(a, b)$$

解析的結果

q の事後分布はベータ分布 $\mathrm{Beta}(a + \sum y_i, b + n - \sum y_i)$

n 局の対局データ $y^n = (y_1, y_2, \ldots, y_n)$ を得た後の，次の対局の予測分布は，$\mathbb{E}[q] = (a + \sum y_i)/(a + b + n)$ をパラメータとするベルヌーイ分布

最尤推定値と事後分布

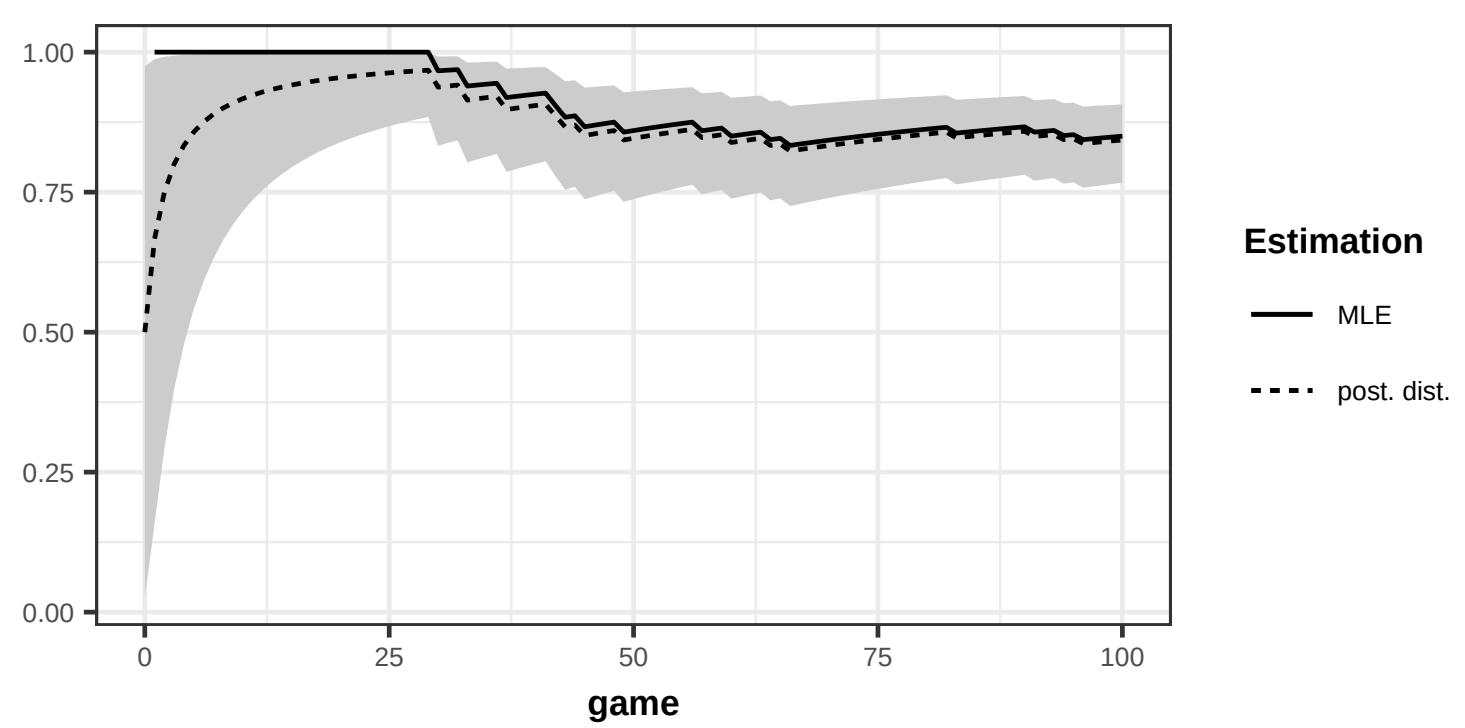

Figure: 推定結果（MLE: 最尤推定値，post dist.: 事後分布平均と95%信頼区間）

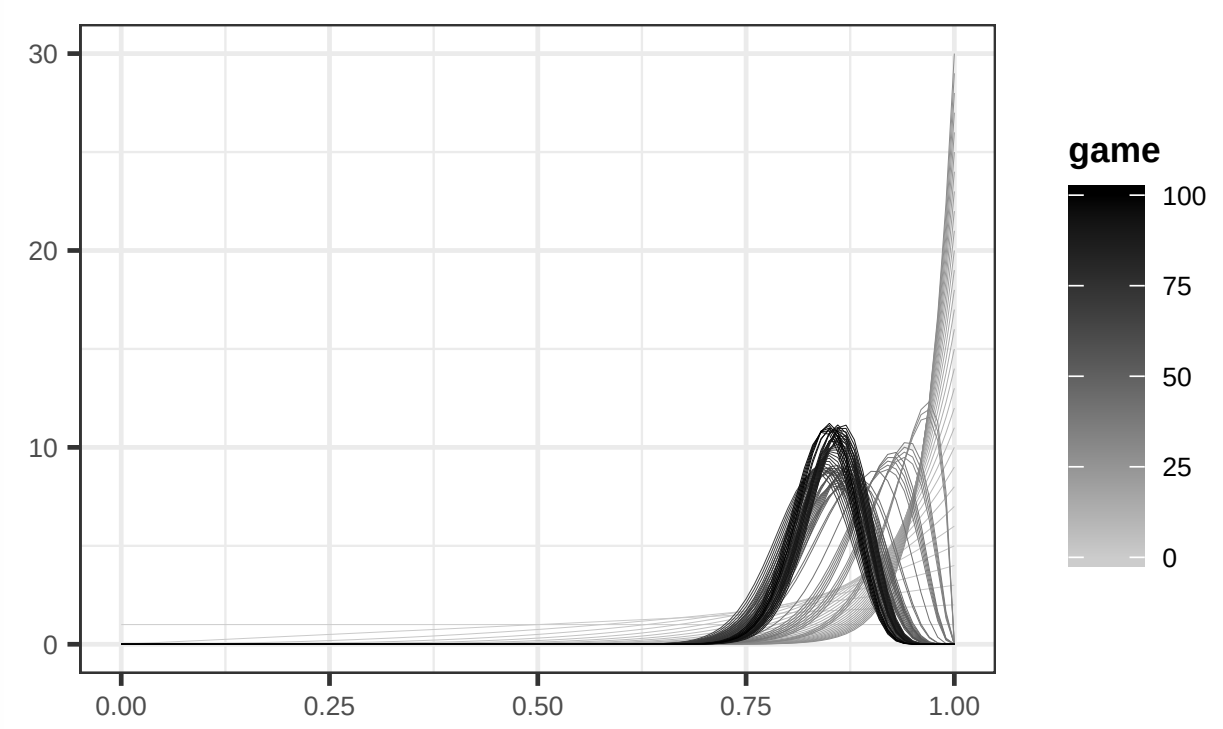

Figure: 事後分布の密度関数の変化

事後分布の平均

事後分布の平均（期待値）を用いた点推定値を **EAP 推定値** (Expected a posterior estimate) という．
一般に，ベータ分布 $X \sim \mathrm{Beta}(a, b)$ の平均は，

$$\mathbb{E}[X] = \frac{a}{a+b}$$

モデル 1 の，データ $y^n = (y_1, y_2, \ldots, y_n)$ を得た後の EAP 推定値

$$\mathbb{E}[q] = \frac{a + \sum y_i}{a+b+n}$$

事後分布の最頻値

事後分布の最頻値を用いた点推定値を **MAP** 推定値 (Maximum a posterior estimate) という．ベータ分布 $X \sim \mathrm{Beta}(a, b)$ の最頻値は，$a > 1, b > 1$ のとき，

$$\mathrm{Mode}[X] = \frac{a-1}{a+b-2}$$

データ $y^n = (y_1, y_2, \ldots, y_n)$ を得た後の MAP 推定値は，

$$\mathrm{Mode}[q] = \frac{a + \sum y_i - 1}{a+b+n-2}$$

- 事前分布を $\mathrm{Beta}(1,1)$（事前情報のない分布）とした場合，MAP 推定値は最尤推定値と等しい．
- 一般に，事前分布をある範囲の一様分布とした場合，MAP 推定値と最尤推定値は等しい．
- なぜなら，事前分布がパラメータについて定数になるために，最尤推定における最大化問題と事後分布の最頻値が等しくなるから．

先手後手で強さが変わるか

i 局目における先手・後手を区別する変数 x_i を導入し，先手なら 1，後手なら 0 と数値をわりあてる．

i 局目の勝敗 Y_i を $x_i q_1 + (1 - x_i) q_0$ をパラメータとするベルヌーイ分布でモデル化する

$$Y_i \sim \mathrm{Bernoulli}(x_i q_1 + (1 - x_i) q_0)$$

Stan code

```
data {
    int N;
    int Y[N];
    int X[N];
}

parameters {
    real<lower=0,upper=1> q1;
    real<lower=0,upper=1> q0;
}

model {
    for (n in 1:N) {
        Y[n] ~ bernoulli(X[n]*q1+(1-X[n])*q0);
        }
}
```

バーンイン 1000，サンプリング 4000，チェーン 4 本
パラメータの事後分布を推定

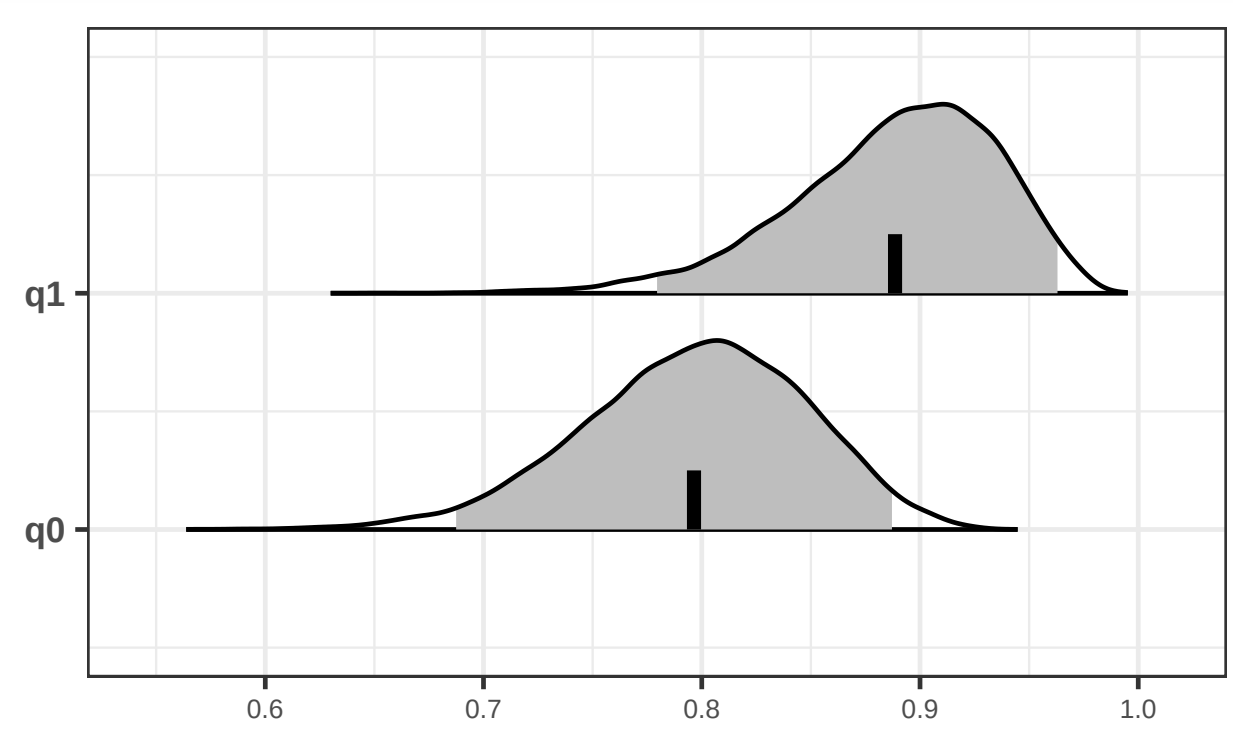

Figure: q_1, q_0 の事後分布（灰色部分が 95%信頼区間）

分布を合成してつくるモデル

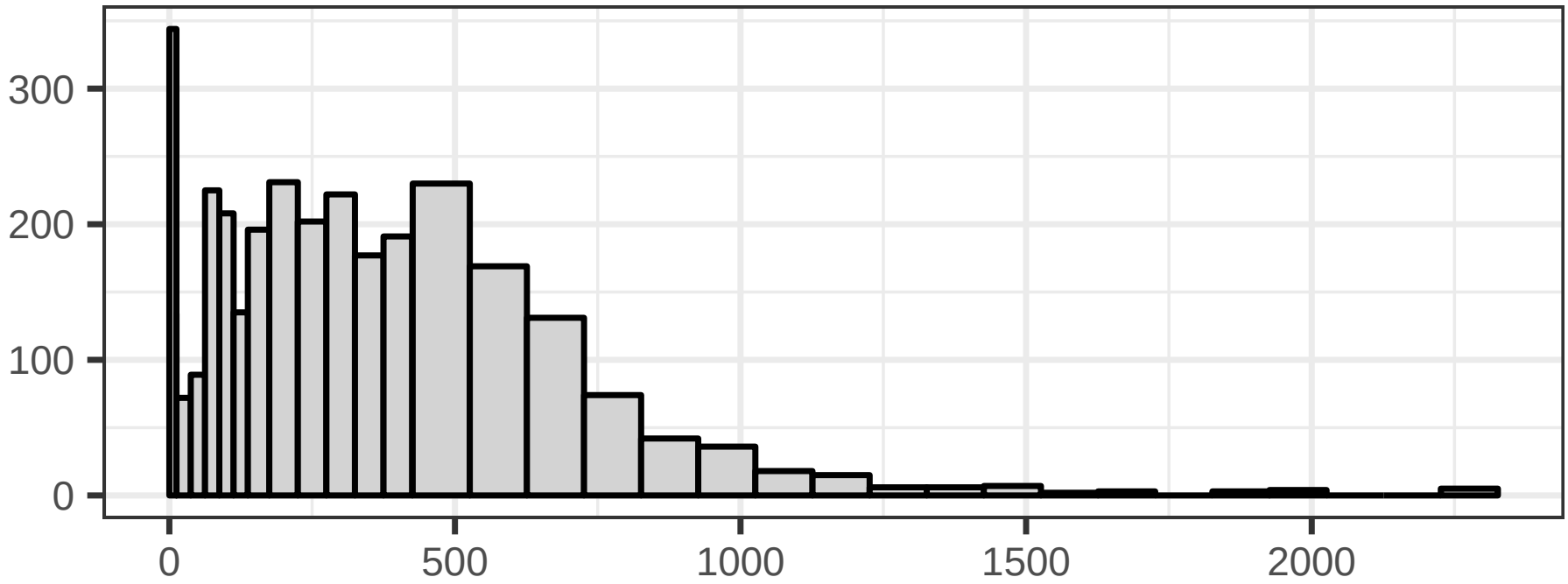

Figure: SSP2015 個人年収（2500 万円以下だけを表示）．横軸は金額（万円），縦軸は人数

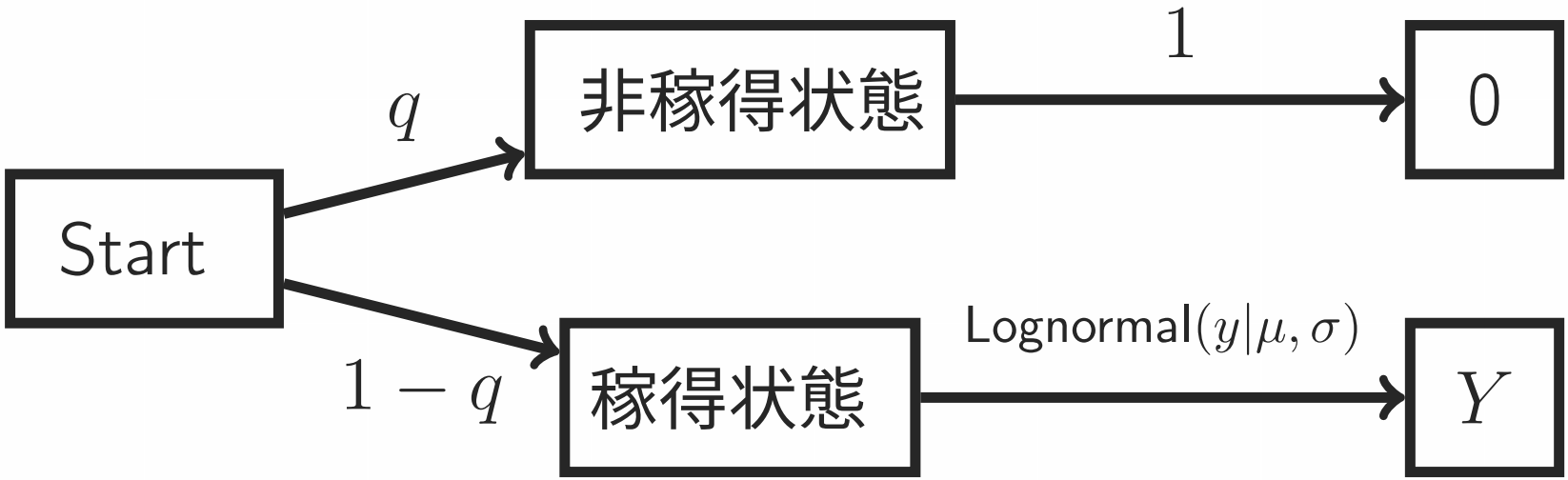

Figure: 分布の合成プロセスを表した樹形図

所得 Y の確率密度関数

$$
\begin{aligned}
HL(y|q,\mu,\sigma) &= \begin{cases} \text{Bernoulli}(1|q), & y=0 \\ \text{Bernoulli}(0|q) \times \text{Lognormal}(y|\mu,\sigma), & y>0 \end{cases} \\
&= \begin{cases} q, & y=0 \\ (1-q)\dfrac{1}{\sqrt{2\pi\sigma^2}y}\exp\left\{-\dfrac{(\log y-\mu)^2}{2\sigma^2}\right\}, & y>0 \end{cases}
\end{aligned}
$$

$HL(y|q,\mu,\sigma)$ はハードル対数正規分布（hurdle lognormal distribution）

$$q_i = \text{logistic}(a_1 + a_2\text{FEM}_i + a_3\text{AGE}_i + a_4\text{EDU}_i)$$
$$\mu_i = b_1 + b_2\text{FEM}_i + b_3\text{AGE}_i + b_4\text{EDU}_i$$
$$Y_i \sim HL(q_i, \mu_i, \sigma) \qquad i = 1, 2, \ldots, n$$

仮定

1. ベルヌーイ分布のパラメータ q（非稼得状態になる確率を表す）が，性別（FEM）や年齢（AGE）や教育年数（EDU）に影響を受ける
2. 稼得状態になった場合の所得のパラメータ μ も，性別や年齢や教育年数に影響を受ける
3. 所得分布は，0 の場合と正の場合とで条件分岐する確率密度関数 $HL(y|q, \mu, \sigma)$ によって定まる

MCMC の結果

確率密度関数をモデル通りに定義

```
1 functions {
2     real HL_lpdf(real Y, real q, real mu, real sigma)
              {
3     if (Y == 0) {
4     return ^^Ibernoulli_lpmf(1 | q);
5     } else {
6     return bernoulli_lpmf(0 | q) + lognormal_lpdf(Y |
            mu, sigma);
7     }
8     }
9 }
```

```
data {
    int N;
    real<lower=0> Y[N];
    int<lower=0> FEM[N];
    real AGE[N]; real EDU[N];
}

parameters {
    real a[4];
    real b[4];
    real<lower=0> sigma;
}
```

決定論的関数は transformed parameters ブロックで

```
1  transformed parameters {
2      real mu[N];
3      real<lower=0,upper=1> q[N];
4      for (n in 1:N){
5      q[n] = inv_logit(a[1]+a[2]*FEM[n]+a[3]*AGE[n]+a
               [4]*EDU[n]);
6      mu[n] = b[1]+b[2]*FEM[n]+b[3]*AGE[n]+b[4]*EDU[n];
7      }
8  }
9
10 model {
11     for (n in 1:N)
12     Y[n] ~ HL(q[n], mu[n], sigma);
13 }
```

バーンイン 1000，サンプリング 1000、チェーン 4

	mean	2.5%	97.5%	n_eff	Rhat
a[1]	-2.730687	-3.77464	-1.73269	1908	1.001
a[2]	1.830781	1.51421	2.16094	3116	0.999
a[3]	-0.015624	-0.02514	-0.00579	3050	1.000
a[4]	0.000363	-0.05662	0.05851	1846	1.000
b[1]	4.617322	4.37122	4.85847	2189	1.001
b[2]	-0.879662	-0.93980	-0.81940	3327	1.000
b[3]	0.007656	0.00507	0.01018	4000	0.999
b[4]	0.072096	0.05842	0.08603	2330	1.001
sigma	0.827043	0.80567	0.84978	3443	1.001
lp__	-20094.427363	-20099.52225	-20091.23233	1782	1.001

モデルが正しいとすれば，非稼得状態になりやすいのは，「女性 (a_2)」「若年 (a_3)」であり，稼得状態になってからは，「男性 (b_2)」「高年齢 (b_3)」「高学歴 (b_4)」であるほど所得が高くなる

よくある GLM

$$
\begin{aligned}
\mu_i &= b_1 + b_2 \mathsf{FEM}_i + b_3 \mathsf{AGE}_i + b_4 \mathsf{EDU}_i \\
Y_i &\sim \mathrm{Lognormal}(\mu_i, \sigma) \\
i &= 1, 2, \ldots, n
\end{aligned}
$$

ハードルモデル : WAIC $= 40198.7 (SE : 204.8)$

GLM : WAIC $= 43840.3 (SE : 143.6)$

どちらのモデルが妥当か

- 単純に WAIC が低いからとか，決定係数が大きいからという機械的選択には意味がない
- 1 回の分析結果からモデルの優劣を決めることはできない．モデル比較はより総合的な視点からおこなうべき
- ハードルモデルを定式化することで，1）所得 0 を非該当扱いにした結果，検出力が下がる，2) 対数化の際に 0 に微少定数を足す等の操作で，推定量にバイアスが生じる，といった不具合を回避できる
- GLM でもそれなりに，データにフィットしたモデルを作ることは可能だが，行為の意味やプロセスをより適切に表現できるのは，現象にあわせてカスタマイズしたモデル

パラメータの生成モデル

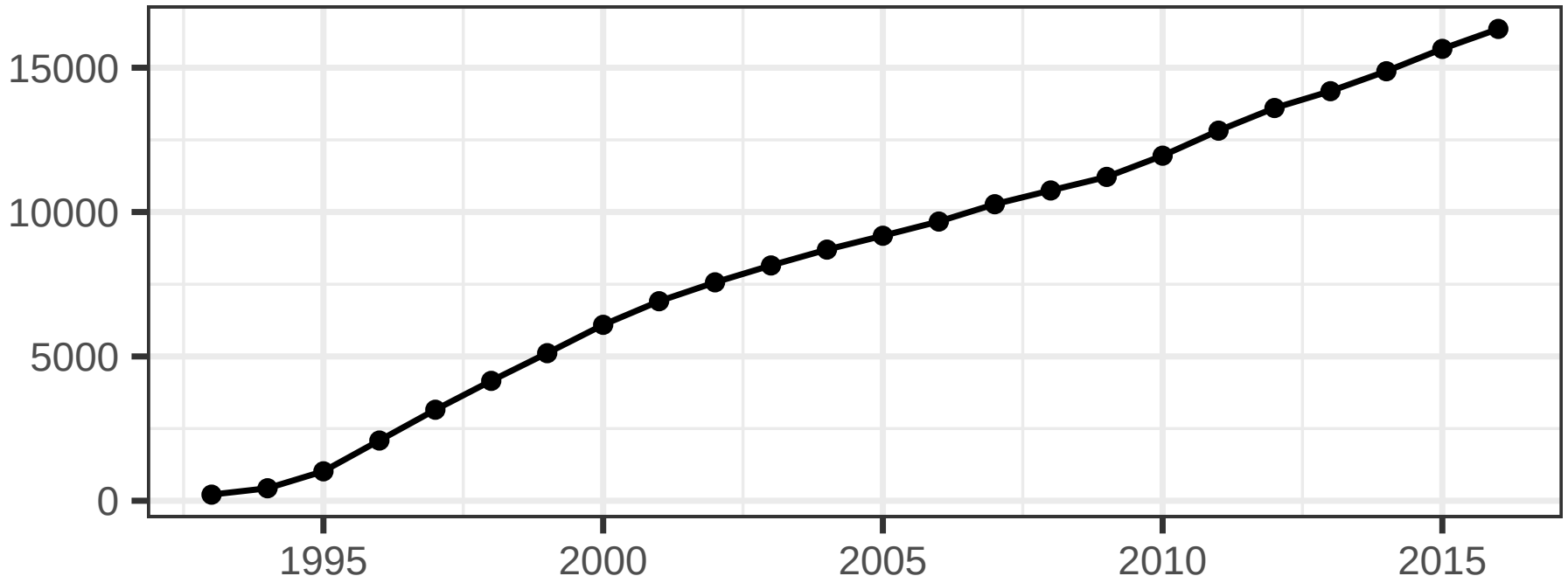

Figure: 携帯電話加入者数（万人）の推移．1993 年〜2017 年．情報通信白書（H16 年版，H23 年版，H30 年版）から作成

関数型を明示的な仮定から導出

1. 契約数 y は時間 t の経過と共に継続的に増加する
2. 契約数には上限がある．これを m とおく
3. 未契約者はランダムに契約者と接触し，未契約者の一部が新たな契約者となる

微分方程式

契約者数 y の瞬間的な増分

$$\frac{dy}{dt} = ky\left(1 - \frac{y}{m}\right)$$

k は増加の早さを決めるパラメータ（$k > 0$）

変数分離型の解

$$\frac{m}{y(m-y)}\frac{dy}{dt} = k \qquad \text{両辺に } \frac{m}{y(m-y)} \text{ をかける}$$

$$\int \frac{m}{y(m-y)}dy = \int k dt \qquad \text{両辺を } t \text{ で積分する}$$

被積分関数を部分分数に分解

$$\frac{m}{y(m-y)} = \frac{a}{y} + \frac{b}{m-y}$$
$$m = a(m-y) + by$$
$$0 \cdot y + m = (b-a)y + am.$$

恒等式を満たす $(b-a) = 0, m = am$ より，$a = 1, b = 1$

$$\int \frac{1}{y} + \frac{1}{m-y} dy = \int k\, dt$$
$$\int \frac{1}{y} dy + \int \frac{1}{m-y} dy = \int k\, dt$$
$$\log y - \log(m-y) = kt + C \quad \text{積分定数を } C \text{ にまとめる}$$
$$\log \frac{y}{m-y} = kt + C$$

初期条件 $t = 0$ のとき $y = y_0$ より積分定数 C を特定

$$\log \frac{y_0}{m - y_0} = k \cdot 0 + C = C.$$

積分定数 $C = \log \frac{y_0}{m-y_0}$ を代入

$$\log \frac{y}{m - y} = kt + \log \frac{y_0}{m - y_0} \qquad \text{積分定数を代入}$$

$$\log \frac{y}{m - y} = \log e^{kt} + \log \frac{y_0}{m - y_0} = \log \frac{y_0 e^{kt}}{m - y_0}$$

$$\frac{y}{m - y} = \frac{y_0 e^{kt}}{m - y_0}$$

$$y = \frac{m y_0}{(m - y_0)e^{-kt} + y_0}. \quad y \text{ について整理}$$

確率モデルに変換

$$Y = \frac{my_0}{(m - y_0)e^{-kt} + y_0} + \varepsilon$$

$$\varepsilon \sim \mathrm{Normal}(0, \sigma)$$

$$Y \sim \mathrm{Normal}\left(\frac{my_0}{(m - y_0)e^{-kt} + y_0}, \sigma\right)$$

使用者数 Y（確率変数）の確率モデルを

$$\text{平均} : \frac{my_0}{(m - y_0)e^{-kt} + y_0}, \text{標準偏差} : \sigma$$

の正規分布で表現．平均パラメータは古典的な微分方程式モデルでヴェアフルスト曲線と呼ばれる関数．

GLM

$$Y = \sum_{i=0}^{k} \beta_i x_i + \varepsilon, \quad \varepsilon \sim \mathrm{Normal}(0, \sigma)$$

β_i の線形結合
誤差項 ε だけが確率変数で正規分布にしたがう

$$Y \sim \mathrm{Normal}\left(\sum_{i=0}^{k} \beta_i x_i, \sigma\right)$$

Stan code

```
parameters {
real <lower=16344,upper=20000>m;
real <lower=0,upper=5>k;
real <lower=0> sigma;
}

transformed parameters{
real mu[n];
for (i in 1:n) mu[i]=(m*y0)/((m-y0)*exp(-k*t[i])+y0);
}

model{
for (i in 1:n) Y[i] ~ normal(mu[i], sigma);
}
```

予測分布

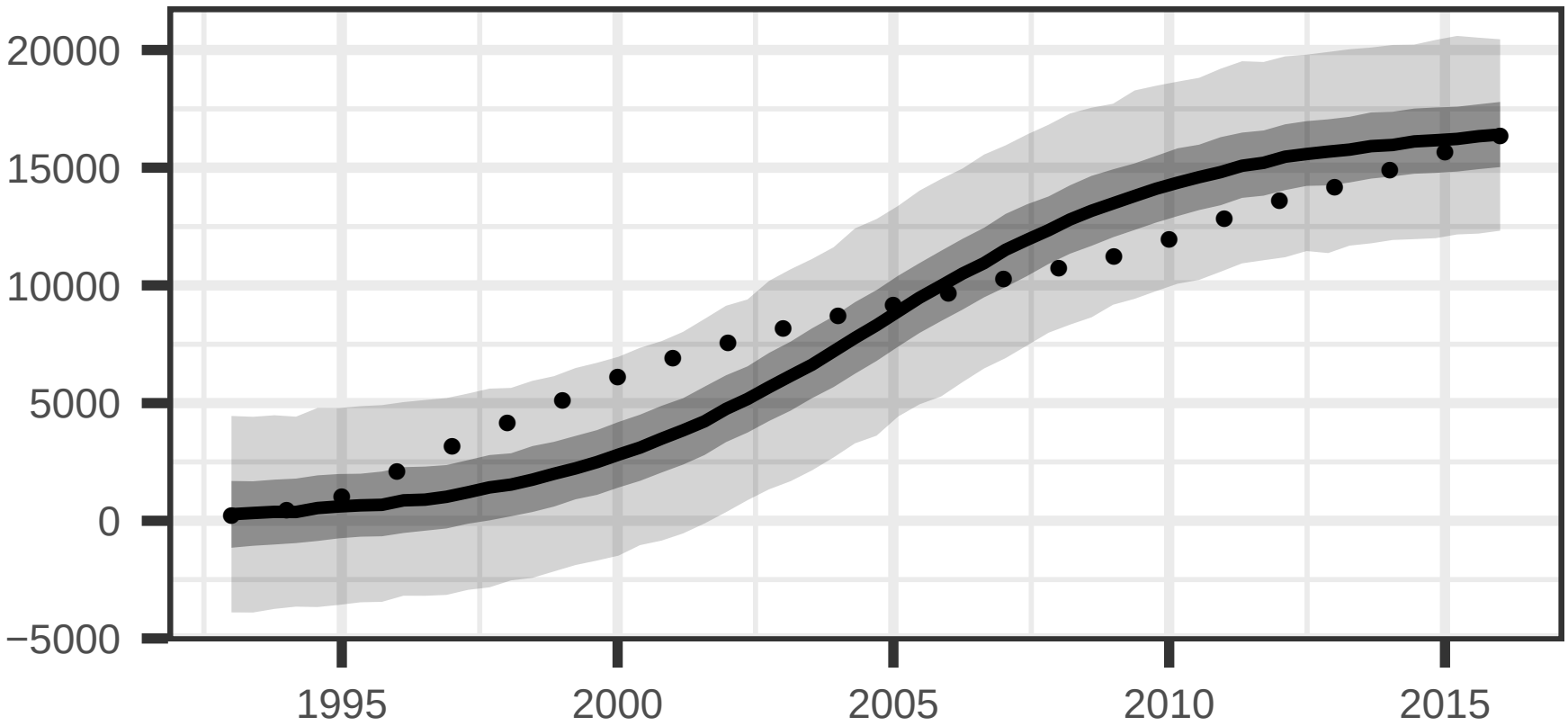

Figure: 観測データとベイズ予測

WAIC 比較

普及プロセスモデル : WAIC $= 434.8(\text{SE} : 4.4)$

GLM : WAIC $= 370.3(\text{SE} : 7.3)$

微分方程式モデルの応用例

- SIR モデルの発展モデル多数
- 広告効果モデル（状態空間モデルに微分方程式モデルを組み込む）. Naik, Prasad A., Murali K. Mantrala and Alan G. Sawyer, 1998, Planning Media Schedules in the Presence of Dynamic Advertising Quality, Marketing Science, Vol. 17, No. 3, pp. 214-235