収入満足度と準拠集団の数理モデル

A Delta Neighborhood Model of Reference Group and Self Evaluation of Economic Status

実践データ駆動科学オンラインセミナー第10回
データ科学・AIにおける数理の威力（2021年9月17日）

東北大学大学院 文学研究科
浜田　宏

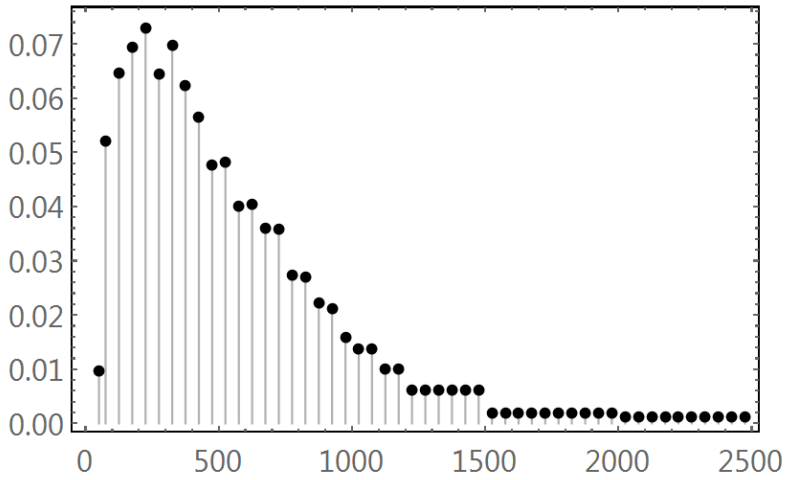
自己紹介

浜田宏（東北大学文学部　教授）

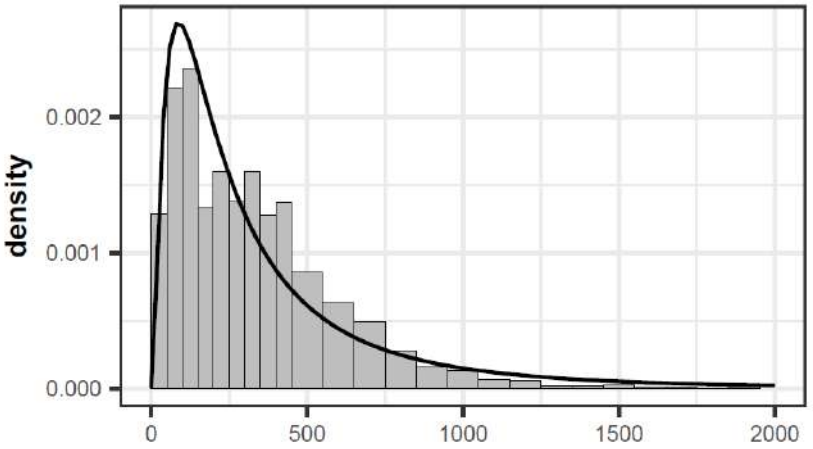
専門：数理社会学

関心：経済的不平等，相対的剥奪

収入の分布

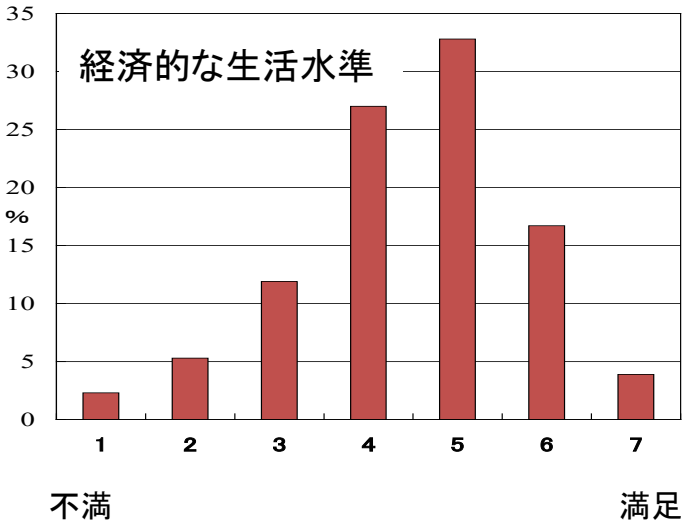


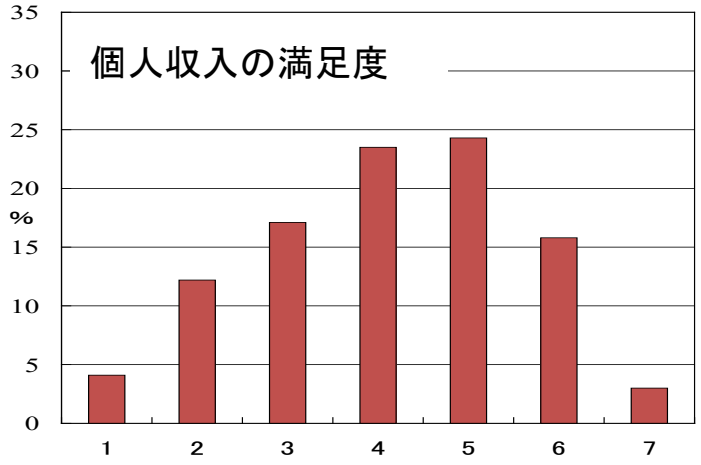
国民生活基礎調査（2016）世帯年収



階層と社会意識全国調査（2015）個人年収.
実線は最尤推定した対数正規分布

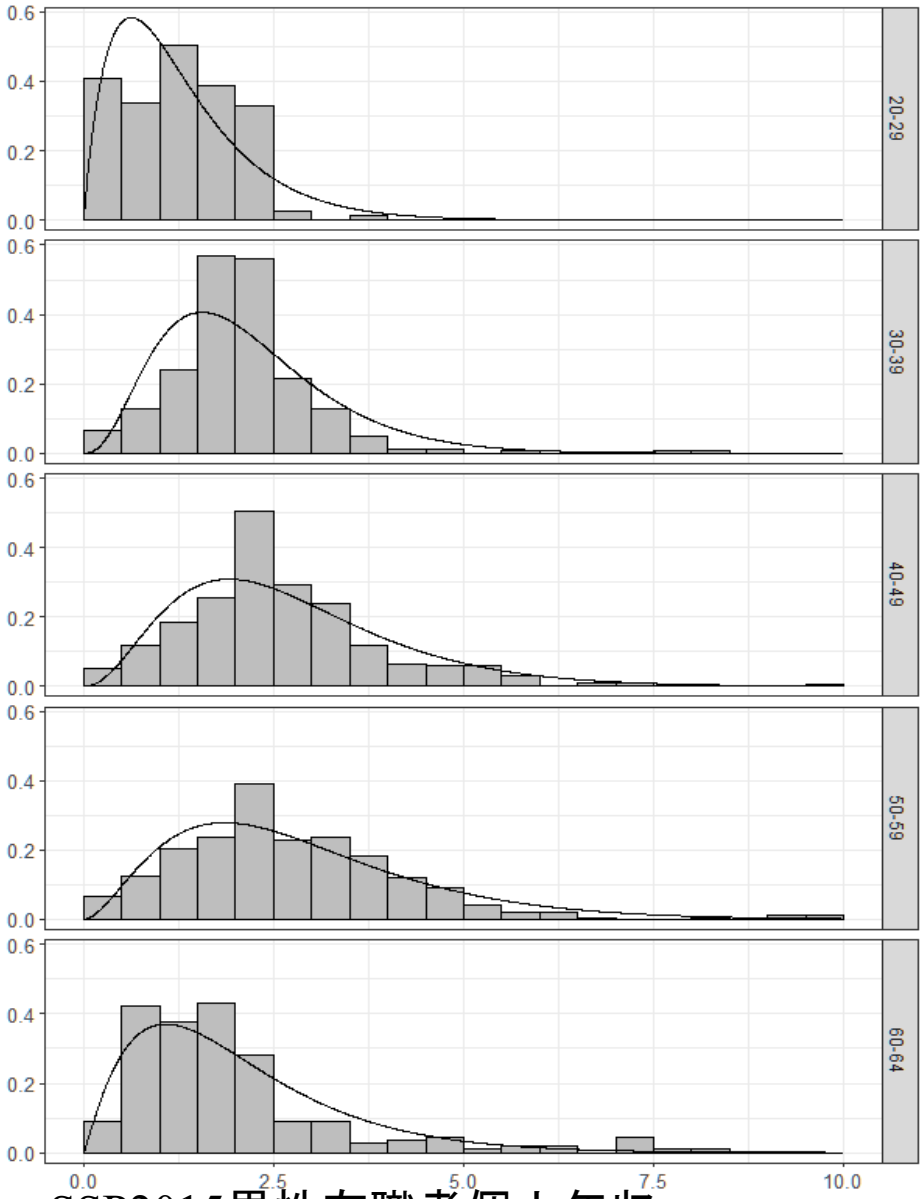
収入満足度の分布



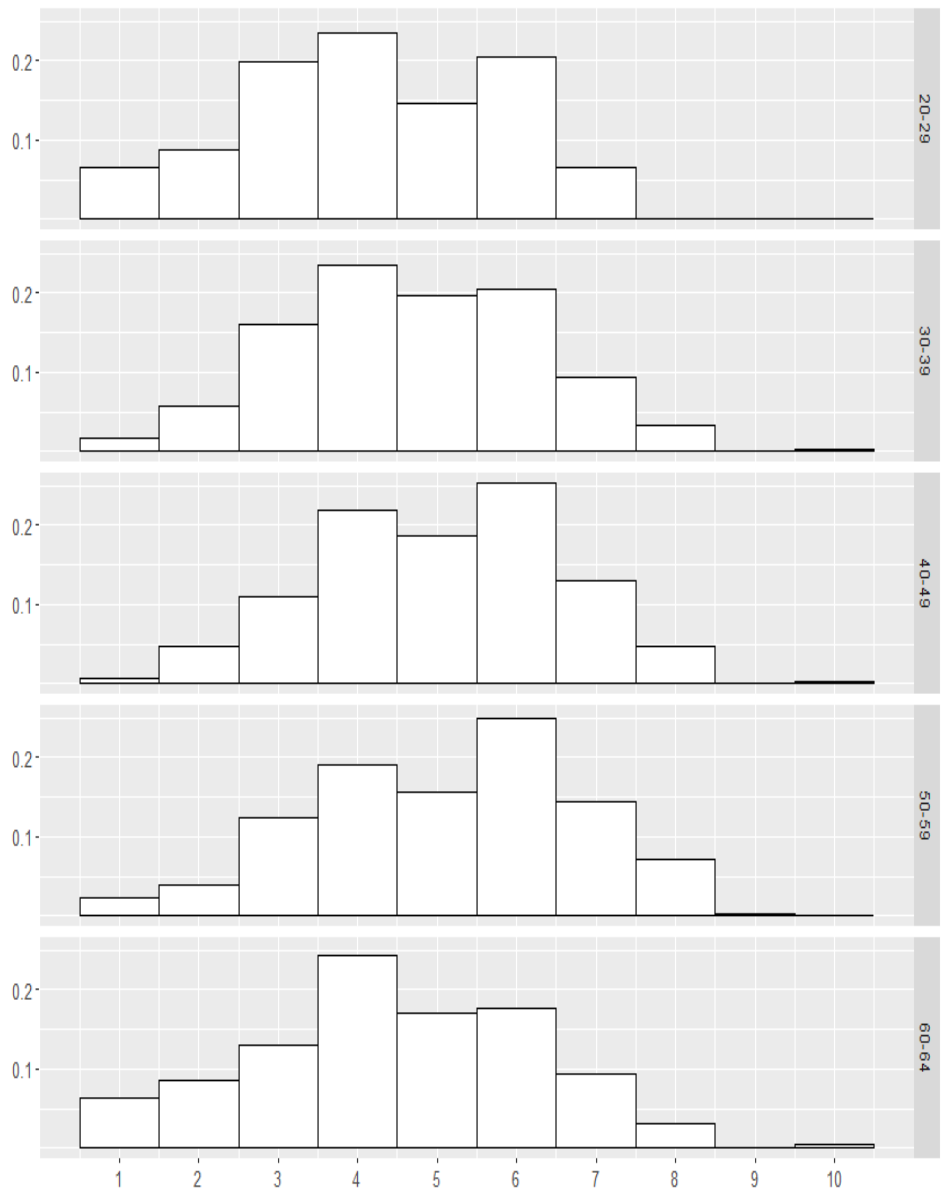


複数の全国調査（JSJP1997,SSM1995, SSP2015）
で同じ傾向

年齢別所得分布



主観的所得水準

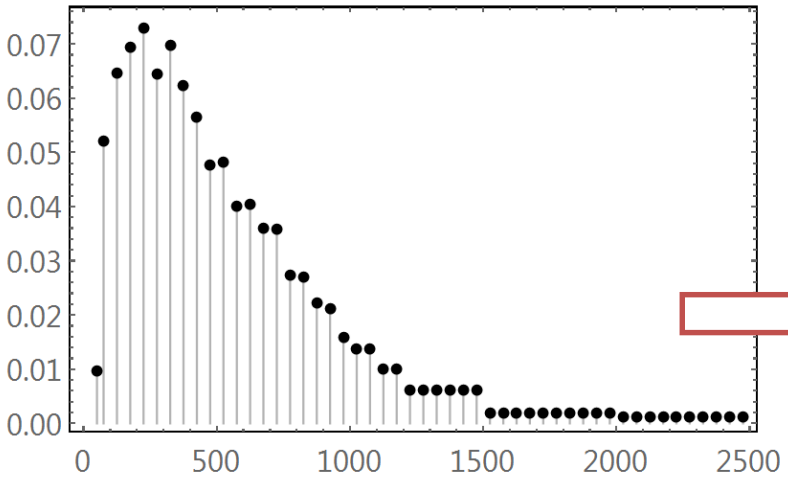


SSP2015男性有職者個人年収

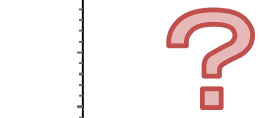
問題：

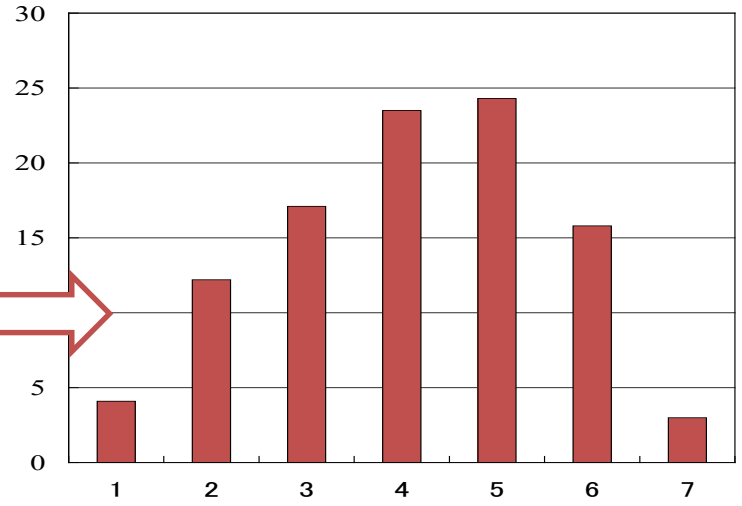
収入の大小で決まるなら、不満はもっと多いはず

なぜ満足するのか？



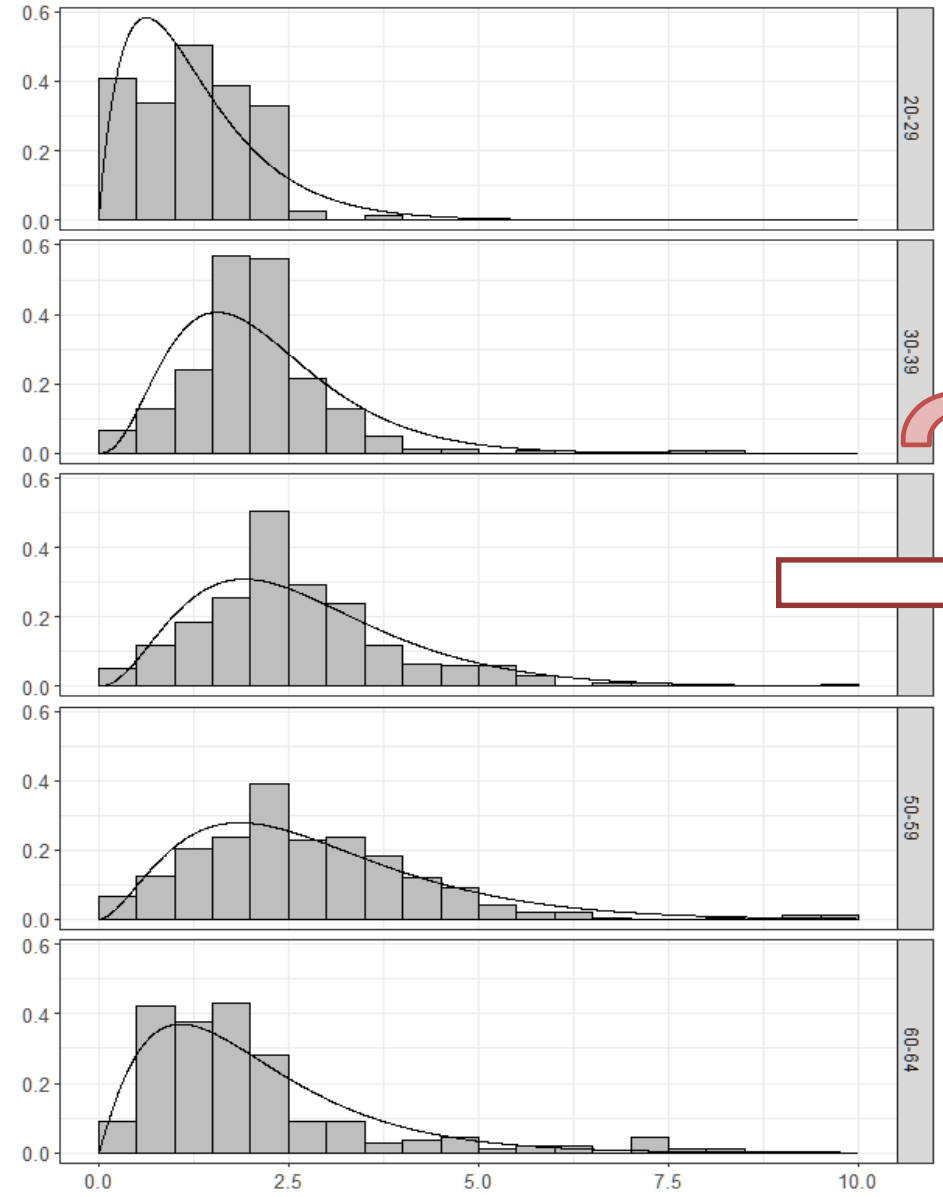
客観的所得分布

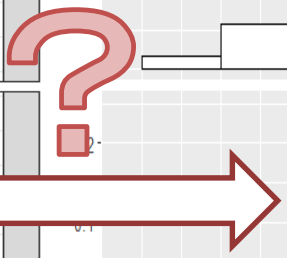




主観的満足度分布

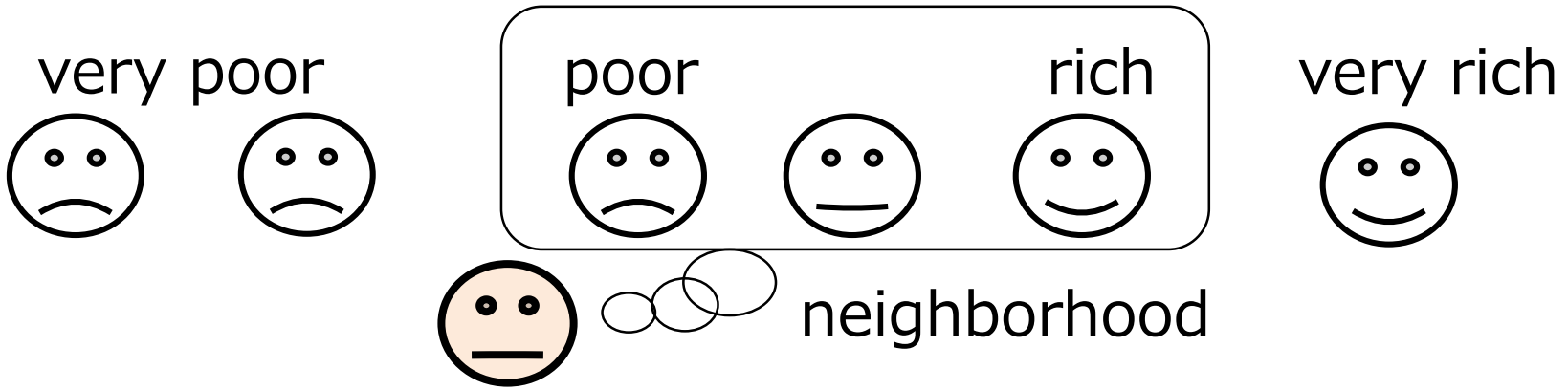
客観的所得分布から主観的満足度分布への変換メカニズムは？



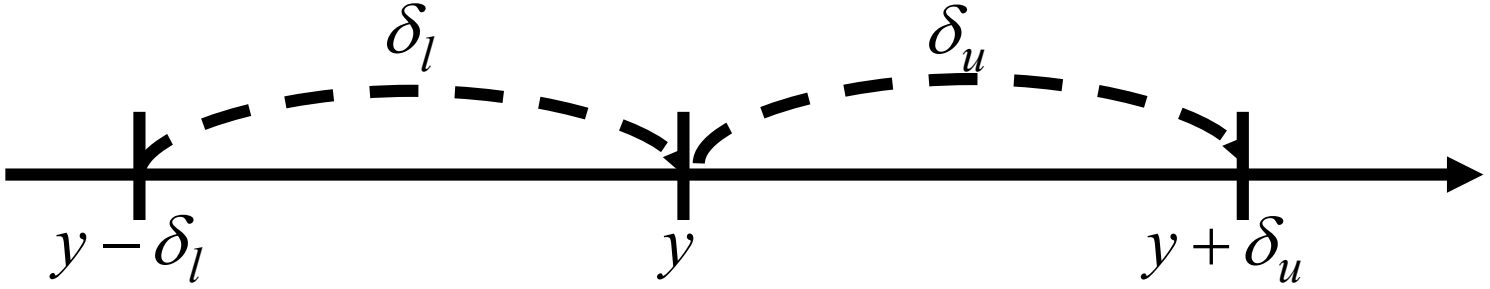


準拠集団理論

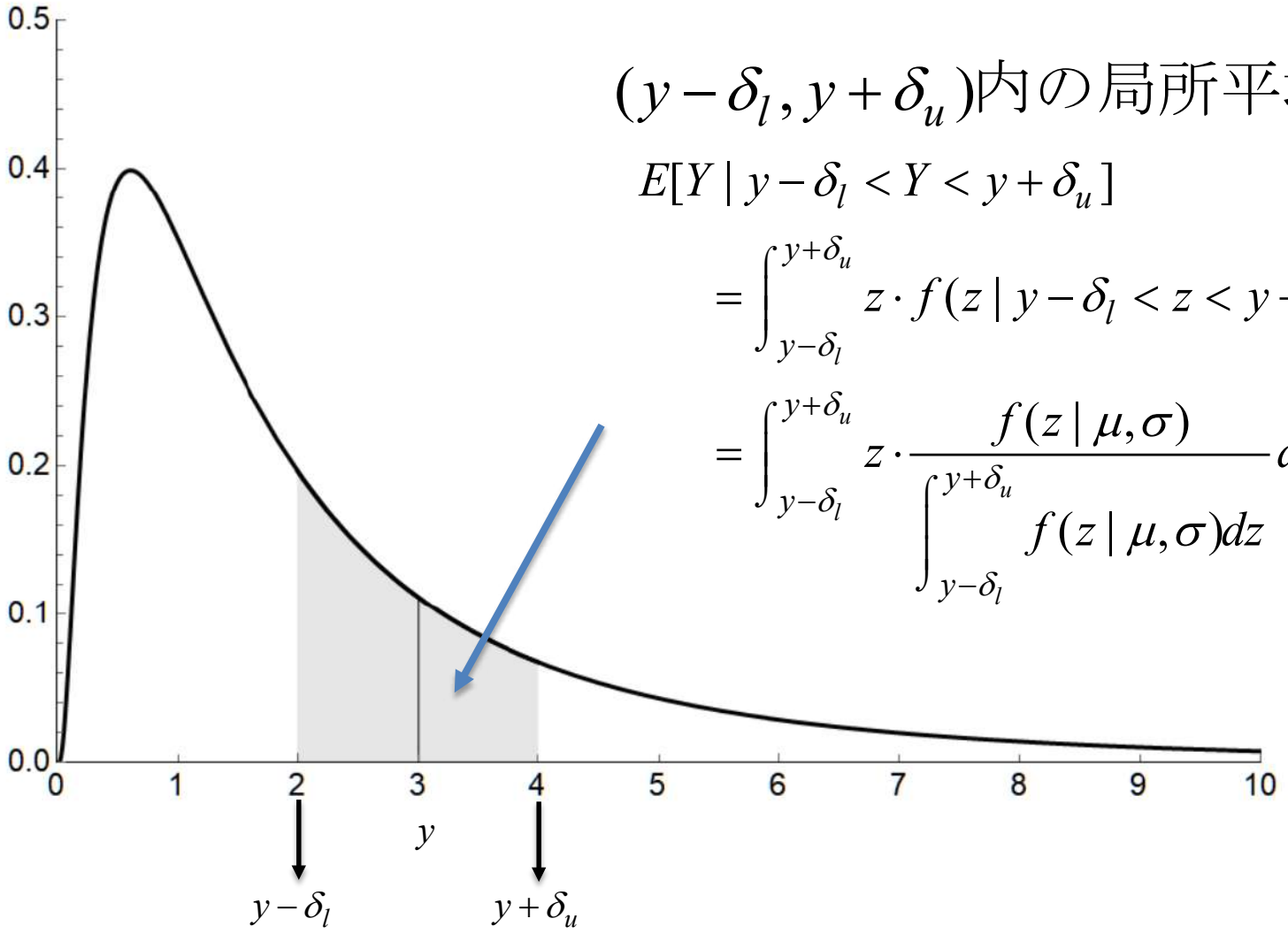
近くの人を比較対象として選ぶ
遠くの人は選ばない



$$(y-\delta_l, y+\delta_u) = \{x \mid y-\delta_l < x < y+\delta_u\} \qquad \delta_l, \delta_u > 0$$



準拠集団の平均所得（y: 本人所得）



$(y-\delta_l, y+\delta_u)$内の局所平均

$$E[Y \mid y-\delta_l < Y < y+\delta_u]$$

$$= \int_{y-\delta_l}^{y+\delta_u} z \cdot f(z \mid y-\delta_l < z < y+\delta_u)\ dz$$

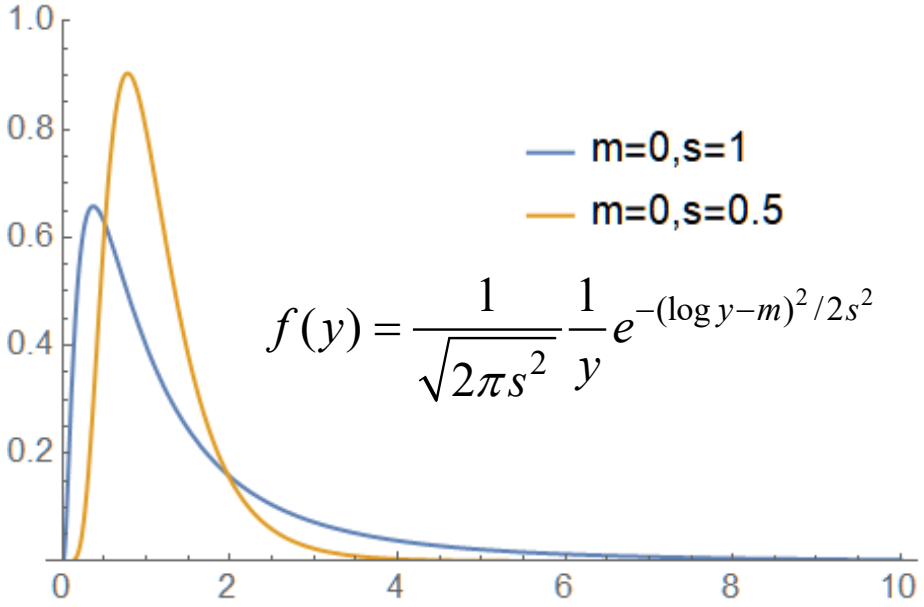
$$= \int_{y-\delta_l}^{y+\delta_u} z \cdot \frac{f(z \mid \mu, \sigma)}{\int_{y-\delta_l}^{y+\delta_u} f(z \mid \mu, \sigma) dz}\, dz$$

Lognormal Distribution

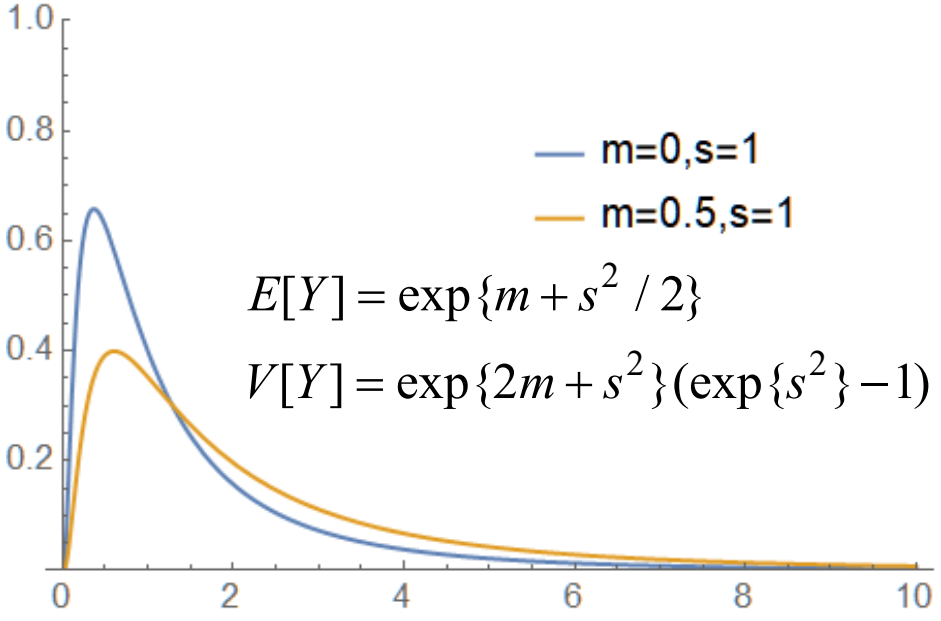
- 対数変換すると正規分布
- 最頻値が左，実現値は正

経験的な意味

- 低・中所得層に集中，超高所得は少数
- $Y \sim \text{Lognormal}(m, s^2) \Rightarrow G = \frac{1}{2\mu} \iint |x - y| f(x) f(y) dxdy = 2\Phi(s/\sqrt{2}) - 1$



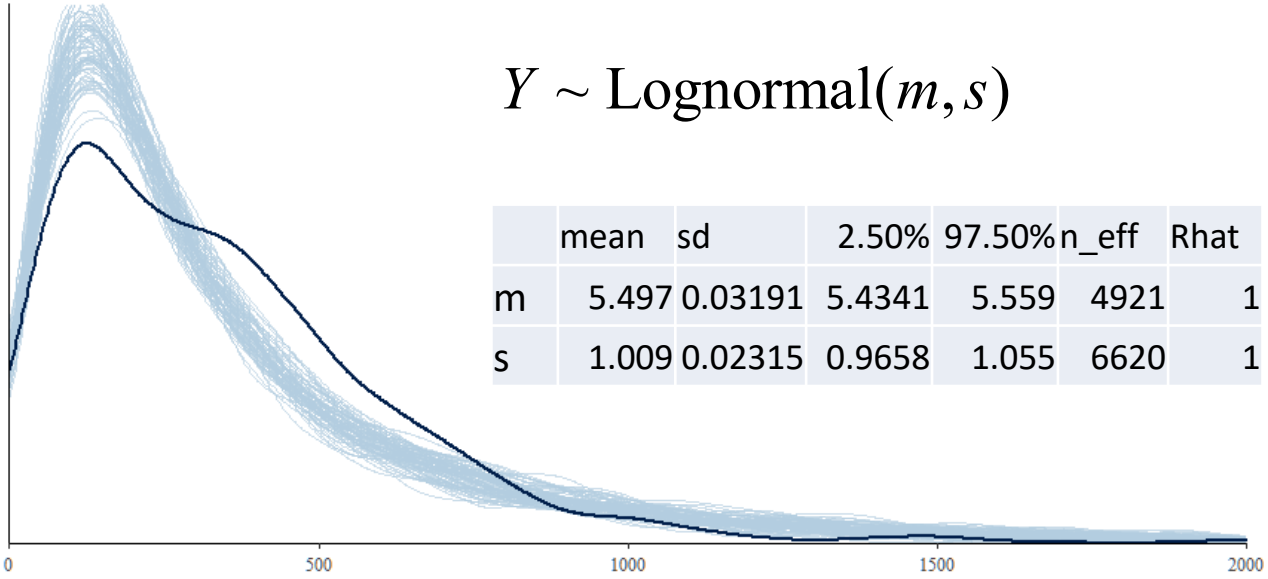
$$f(y) = \frac{1}{\sqrt{2\pi s^2}} \frac{1}{y} e^{-(\log y - m)^2 / 2s^2}$$



$$E[Y] = \exp\{m + s^2/2\}$$

$$V[Y] = \exp\{2m + s^2\}(\exp\{s^2\} - 1)$$

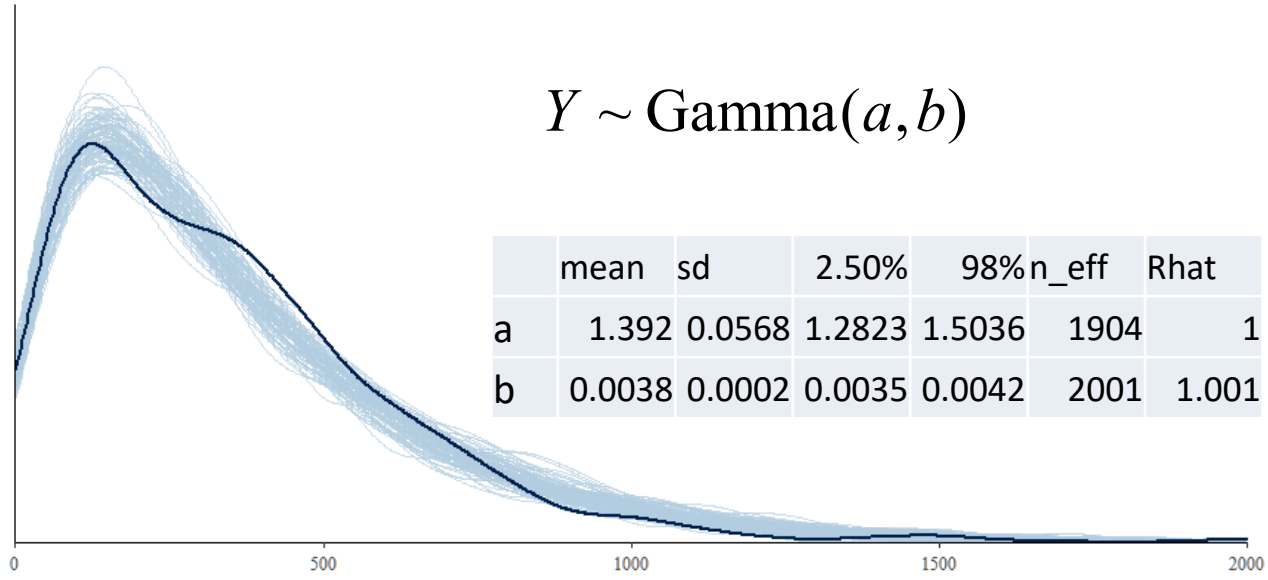
所得分布の確率モデル比較



$$Y \sim \mathrm{Lognormal}(m, s)$$

	mean	sd	2.50%	97.50%	n_eff	Rhat
m	5.497	0.03191	5.4341	5.559	4921	1
s	1.009	0.02315	0.9658	1.055	6620	1

	Estimate	SE
elpd_waic	-6924.5	30
p_waic	2.6	0.2
waic	13849	59.9



$$Y \sim \mathrm{Gamma}(a, b)$$

	mean	sd	2.50%	98%	n_eff	Rhat
a	1.392	0.0568	1.2823	1.5036	1904	1
b	0.0038	0.0002	0.0035	0.0042	2001	1.001

	Estimate	SE
elpd_waic	-6869.8	33.6
p_waic	2.6	0.5
waic	13739.5	67.2

個人年収SSP2015

ガンマ分布の条件付き確率密度

$$E[Y \mid y-\delta_l < Y < y+\delta_u]$$

$$= \int_{y-\delta_l}^{y+\delta_u} z \cdot \frac{f(z \mid \mu,\sigma)}{\int_{y-\delta_l}^{y+\delta_u} f(z \mid \mu,\sigma)dz} dz$$

準拠集団平均所得

$$PDF : f(y \mid a,b) = \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by}, \quad y > 0$$

PDFにガンマ分布を仮定

$$\int_c^d f(z \mid a,b)\, dz = \frac{1}{\Gamma(a)} \{\Gamma(a,cb) - \Gamma(a,db)\}$$

条件付き確率密度

where $\Gamma(a,z) = \int_z^\infty t^{a-1} e^t \ dt$

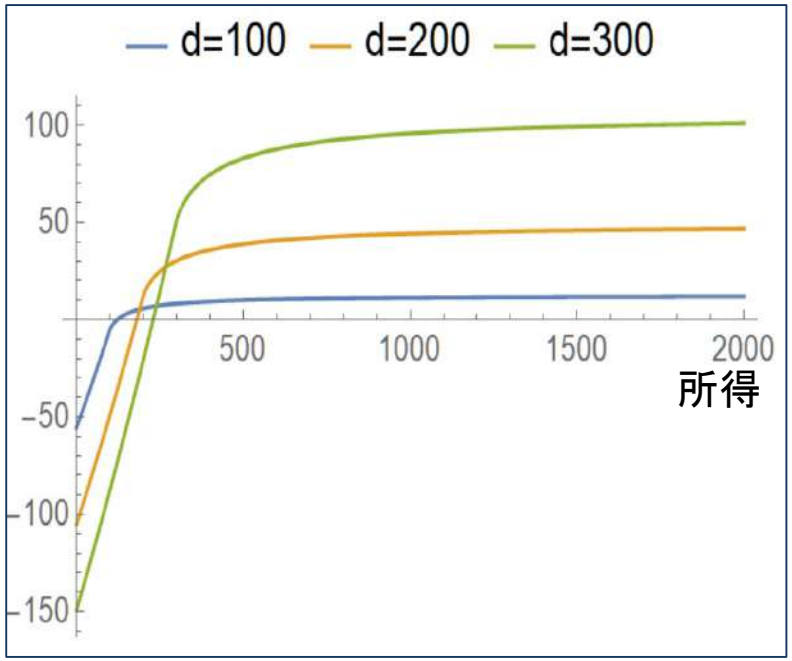
Incomplete Gamma function

準拠集団平均所得を不完全ガンマ関数で近似できる

限界効用の逓減

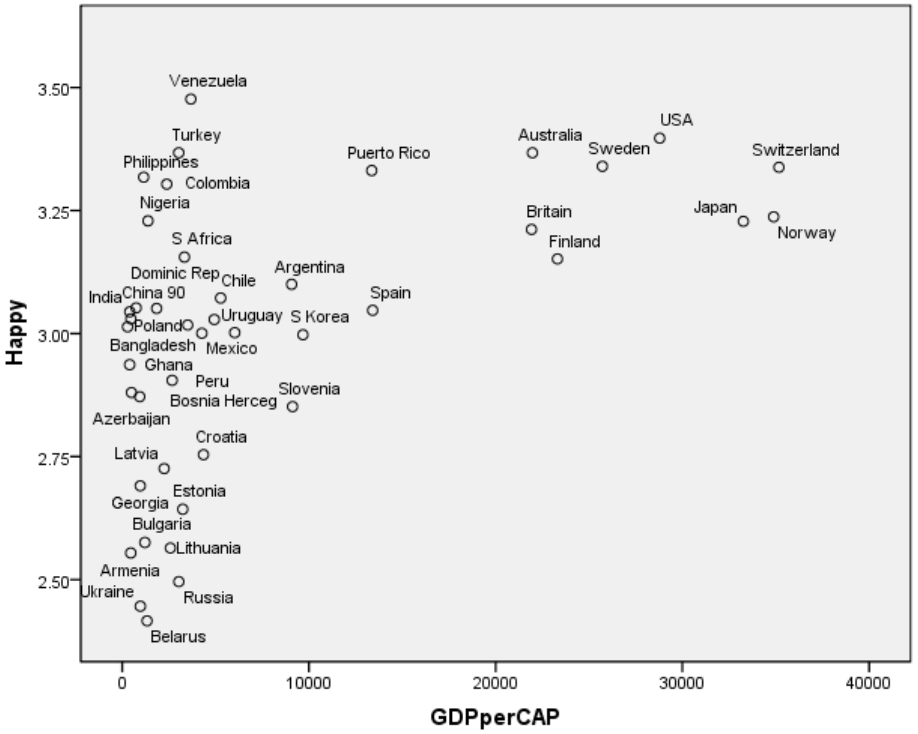
Model

効用=収入－他者平均



Data

主観的幸福感



各国の一人あたりGDP（US$ 購買力平価）と幸福感の関係

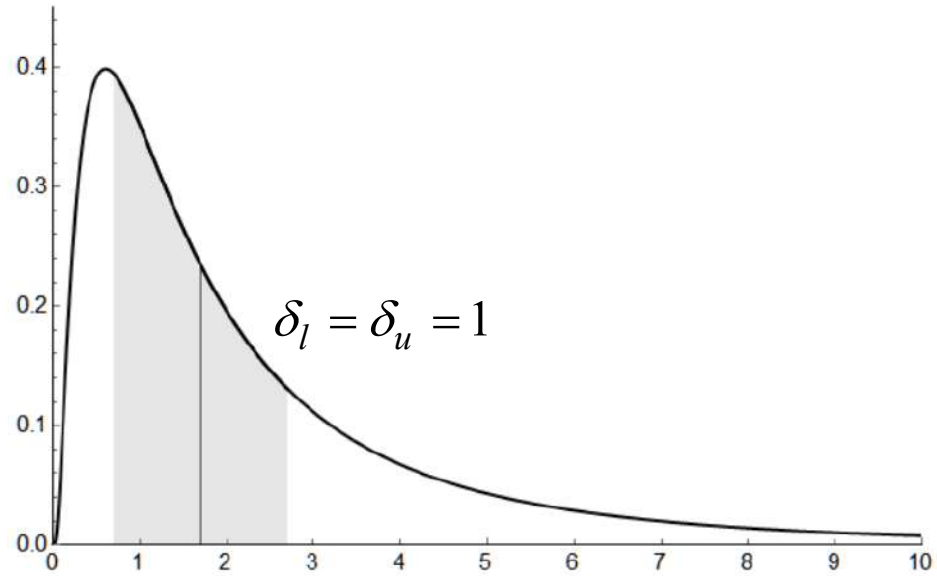
World Value Survey 2000, UNDP report

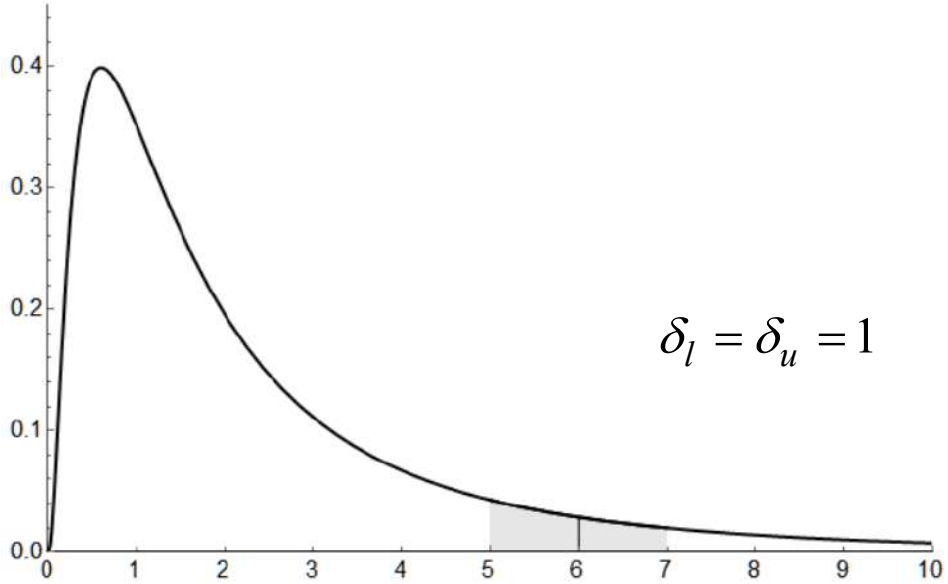
$$y - f(y, \delta_l, \delta_u) \quad \text{where} \quad Y \sim \text{Gamma}(a, b)$$

$$f(y, \delta_l, \delta_u)$$

$$= \frac{\Gamma(a+1, b(y-\delta_u)) - \Gamma(a+1, b(y+\delta_u))}{b\{\Gamma(a, b(y-\delta_l)) - \Gamma(a, b(y+\delta_u)\}}$$

限界効用逓減の直感的しくみ





中層にいても高層にいても相対的に満足.
ただし自分と準拠集団との差はあまり変わらない.

収入が増えると満足だが効用増加は鈍くなる
低い収入でもそれなりに満足

⇒2つの傾向を矛盾なく説明

年齢の影響

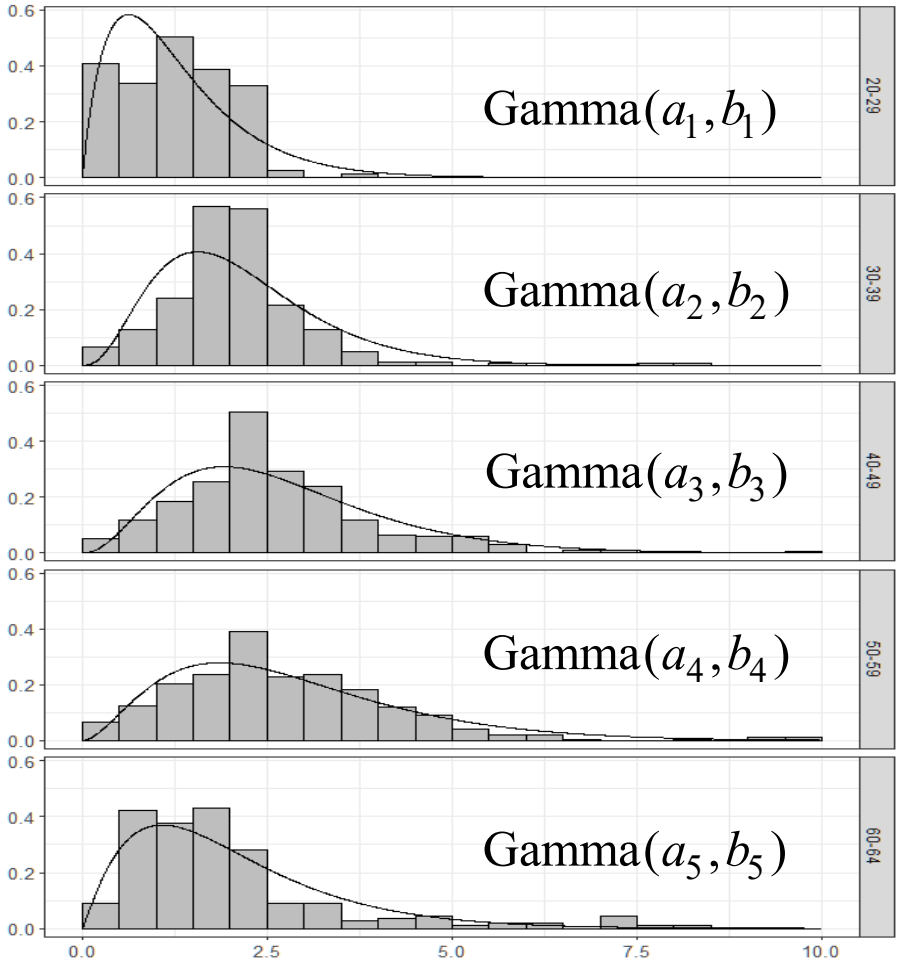
収入を比べる他者との共通点（飯田 2009）

	2008年	2009年
年齢	36.00%	49.30%
所得水準	6.10%	8.20%
性別	8.40%	6.40%
職業	26.70%	16.80%
学歴・出身校	7.00%	4.60%
住んでいる地域	6.70%	7.10%
その他	1.70%	2.10%
共通点はない	7.30%	5.40%
合計	100.00%	100.00%
	n＝344	n＝280

年齢が近いほど比較対象になりやすい

職業が同じなら比較対象になりやすい

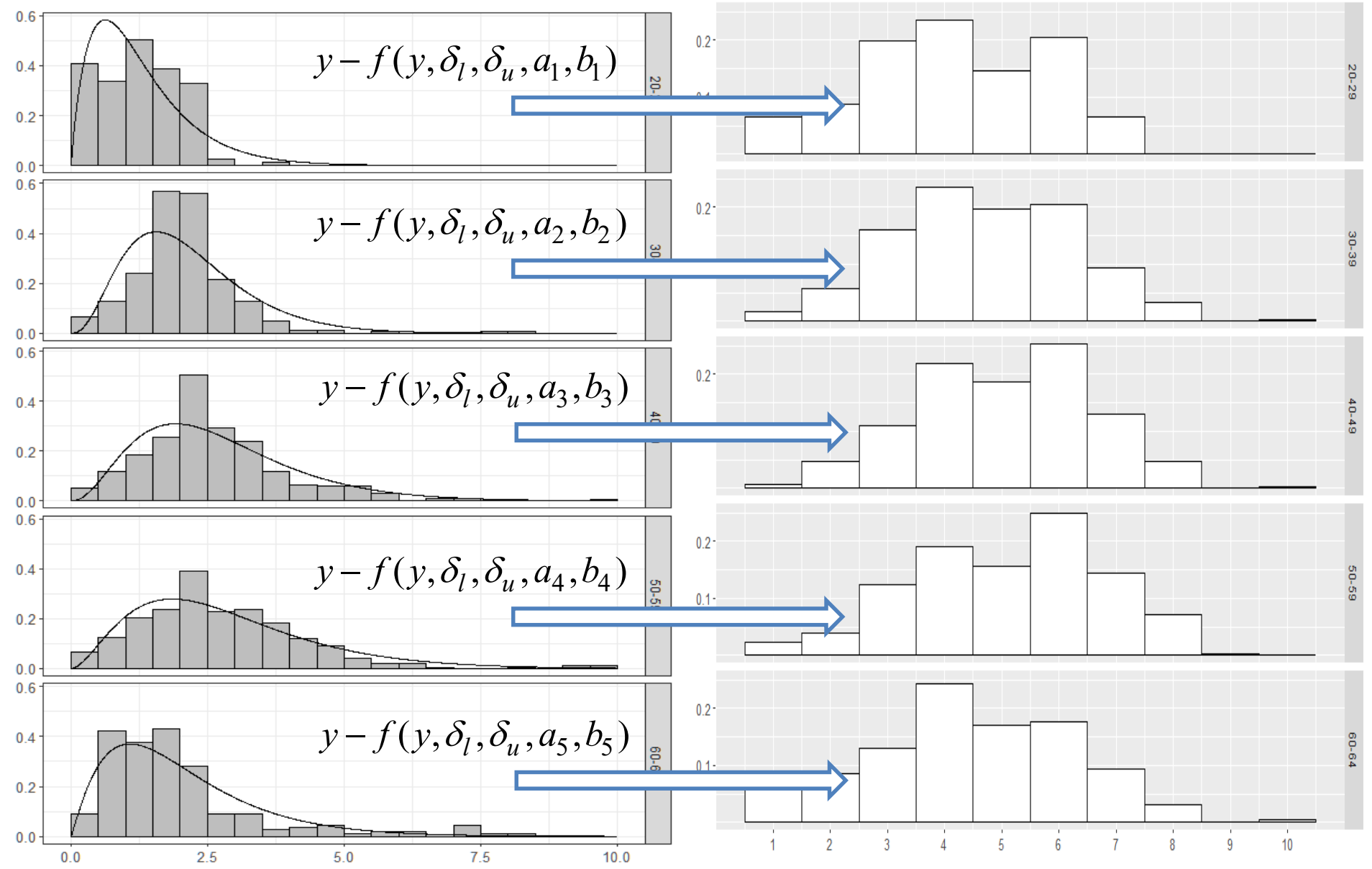
```
parameters {
  real <lower=0>a[5];  real <lower=0>b[5];
}
model {
  for( i in 1 : N ) {
    target += gamma_lpdf(X[i] |a[age[i]], b[age[i]]);  }
}
```



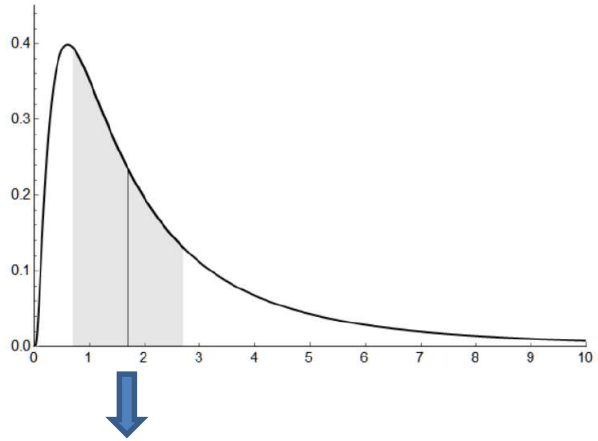
年齢別に所得分布パラメータを推定
（事前分布が一様分布なので事後分布の平均値は最尤推定値）

	mean	sd	2.50%	97.50%	n_eff	Rhat
a[1]	1.988	0.196	1.631	2.395	5281.217	1.000
a[2]	3.681	0.287	3.137	4.255	5885.100	1.000
a[3]	3.355	0.243	2.897	3.850	5781.154	1.000
a[4]	2.839	0.200	2.459	3.238	6064.448	1.000
a[5]	2.159	0.190	1.802	2.552	5756.719	1.000
b[1]	1.578	0.178	1.250	1.950	5413.854	1.000
b[2]	1.719	0.144	1.450	2.008	5816.238	1.000
b[3]	1.230	0.096	1.053	1.422	5831.241	1.000
b[4]	0.991	0.077	0.846	1.146	6074.977	1.000
b[5]	1.071	0.106	0.871	1.287	5524.442	1.000

客観的所得分布から主観的満足度分布への変換プロセス



準拠集団比較モデルをベイズ推定



準拠集団の平均をガンマ分布の局所平均で近似

比較の範囲δ_u, δ_lは事後分布を計算する

$$m[i] = \frac{\Gamma(a_k + 1, b_k c_k) - \Gamma(a_k + 1, b_k d_k)}{b\{\Gamma(a_k, b_k c_k) - \Gamma(a_k, b_k d_k)\}}, \quad \text{年代} k = 1,2,3,4,5$$

$$c_k = y[i] - \delta_{lk}, \quad d_k = y[i] + \delta_{uk},$$

$$\delta_{lk} \sim \text{Uniform}(0,20), \delta_{uk} \sim \text{Uniform}(0,20)$$

$$w[i] \sim N(\beta_0 + \beta_1 (y[i] - m[i]) + \sum_j \beta_j x_j[i], \sigma^2) \quad \text{個人} i = 1,2,\ldots,n$$

$$y : \text{income}, m : \text{relative income}, x_j : \text{ex var}$$

$$w : \text{economic well-being}$$

事後分布

定義 (事後分布)

パラメータ θ の事後分布（posterior distribution）を

$$p(\theta|x^n) = \frac{p(x^n|\theta)\varphi(\theta)}{\int p(x^n|\theta)\varphi(\theta)d\theta}$$

と定義する．$p(x^n|\theta) = \prod_{i=1}^n p(x_i|\theta)$．積分の範囲 S はパラメータ θ のとりうる範囲（$\theta \in S$）とする．事後分布の分母（正規化項）を周辺尤度 (marginal likelihood) という

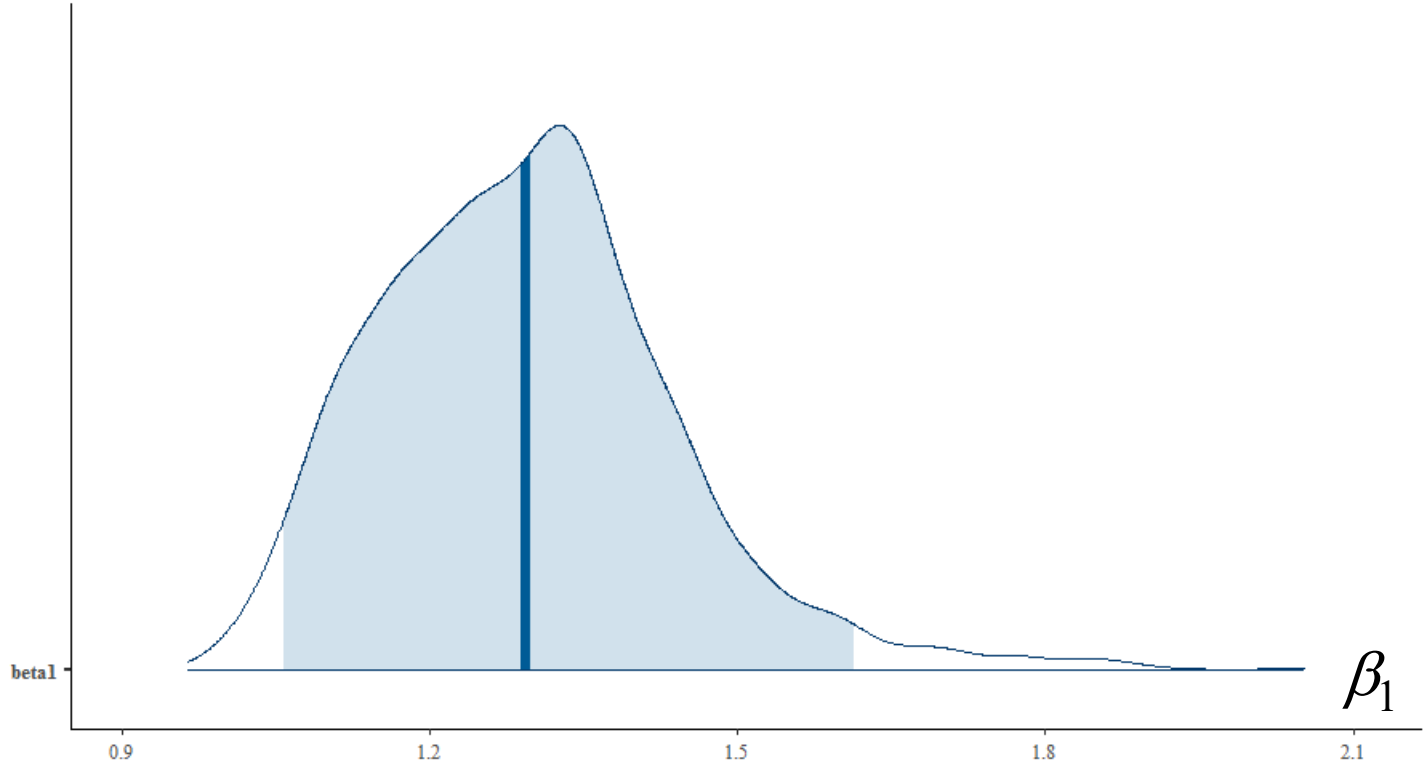
事後分布の直感的意味：データを条件とする，パラメータの条件付き確率

周辺尤度の直感的意味:パラメータの事前分布を考慮したサンプルの同時確率関数

$$W[i] \sim N(\beta_0 + \beta_1(Y[i] - E[Y \mid Y[i] - \delta_l < Y < Y[i] + \delta_u]), \sigma)$$

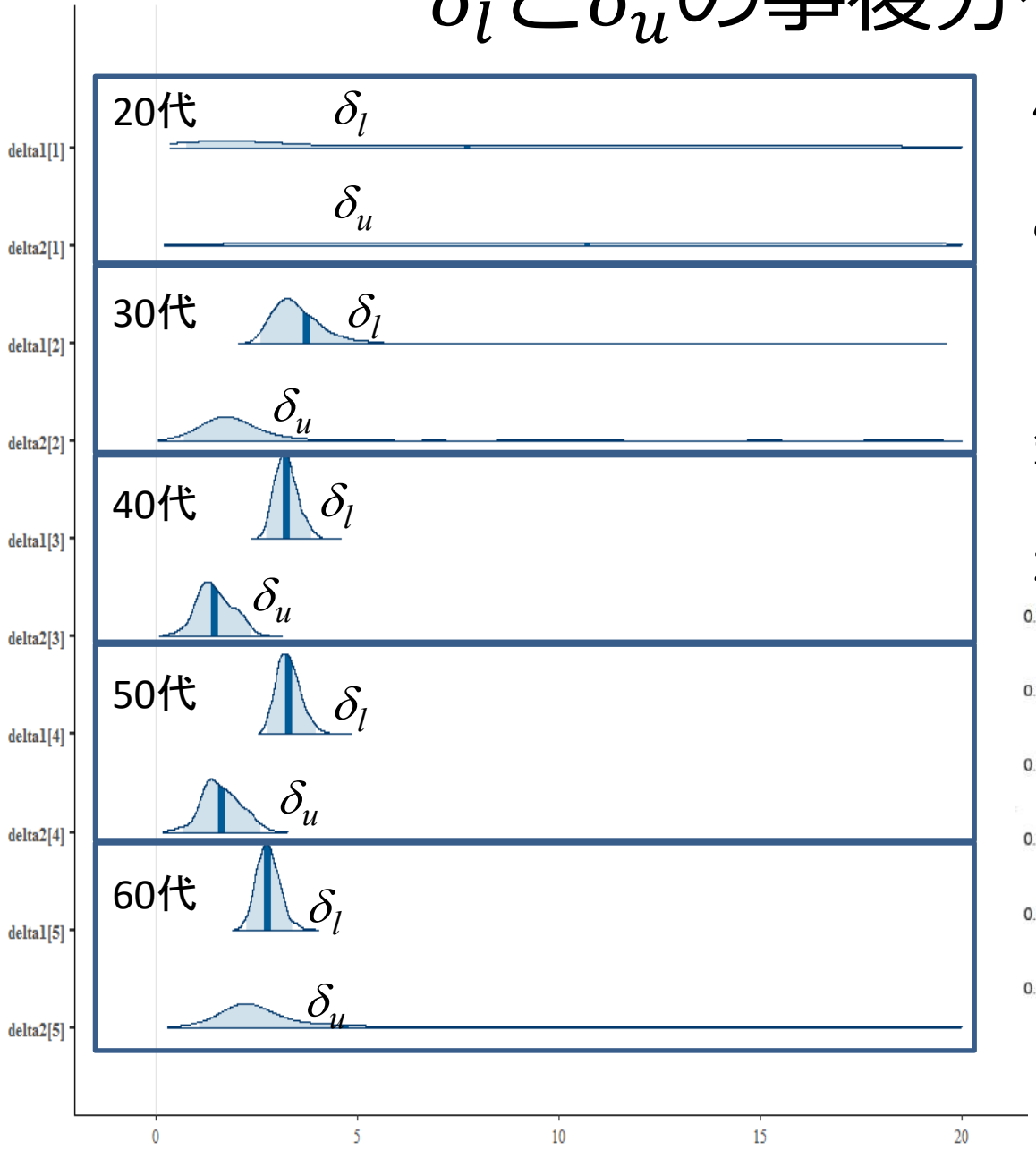
仮説：自分の収入＞準拠集団の収入なら満足．逆なら不満

結果：理論モデルの予想と一致



β_1の事後分布

δ_lとδ_uの事後分布比較

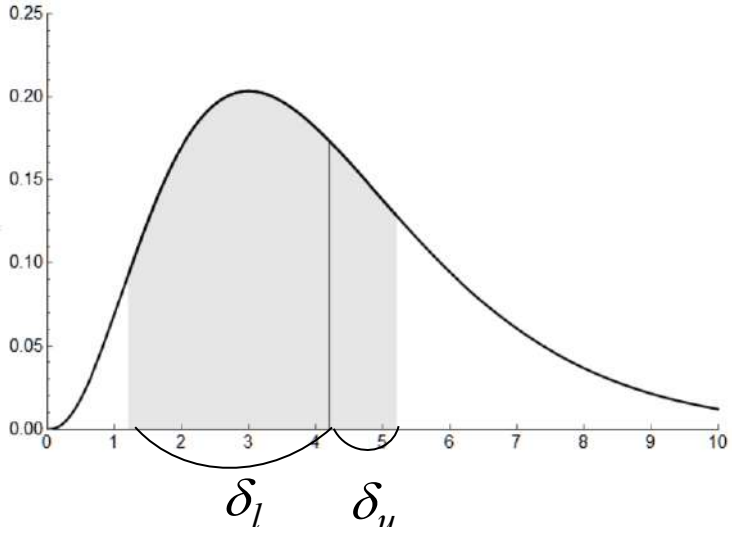


40代と50代は安定

δ_lの平均がδ_uの平均より大きい

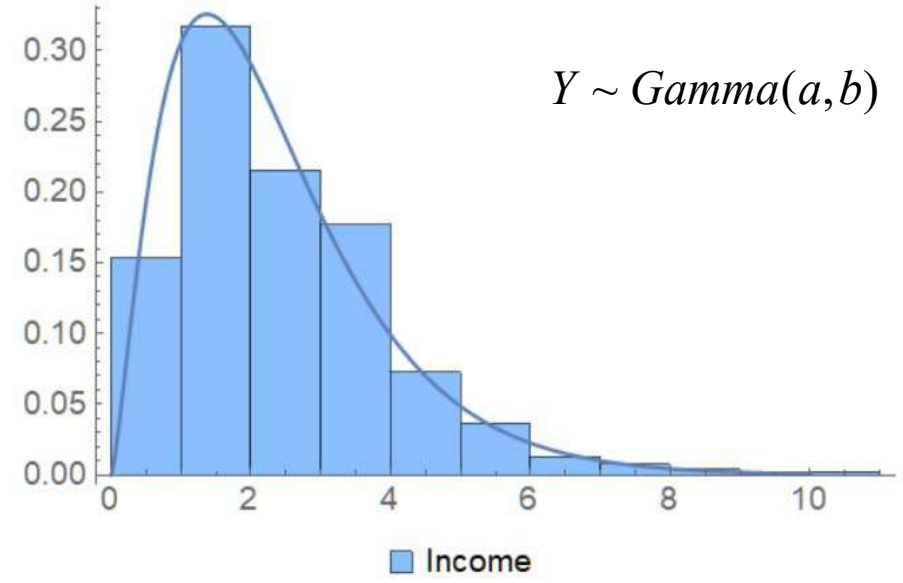
「下」をみる範囲が「上」をみる範囲より広い可能性を示唆

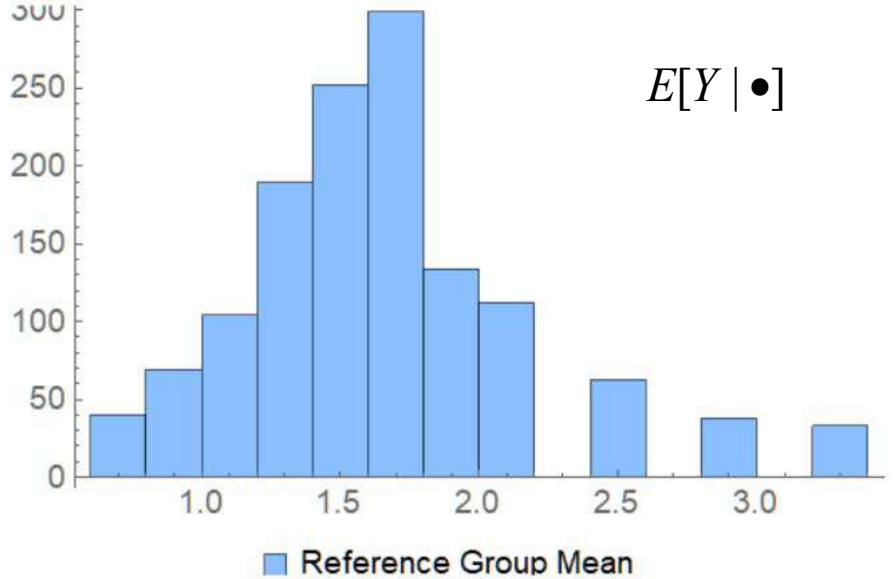
20代と60代は傾向不明

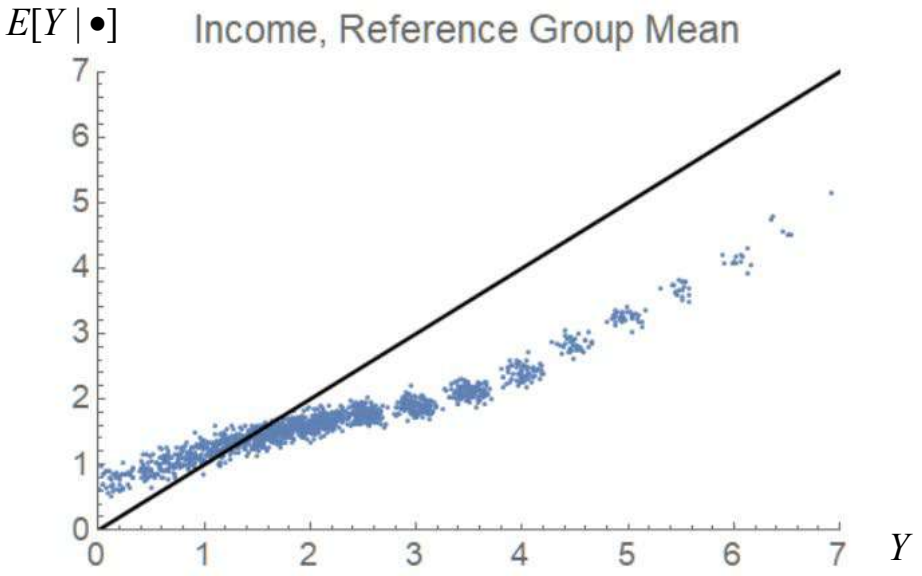


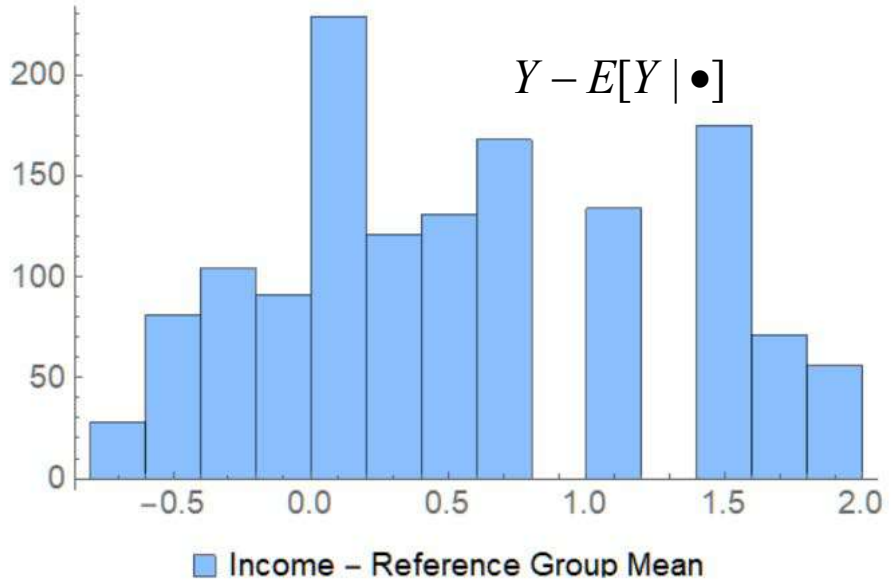
40代所得分布と準拠集団範囲（平均）

Income and Reference Group







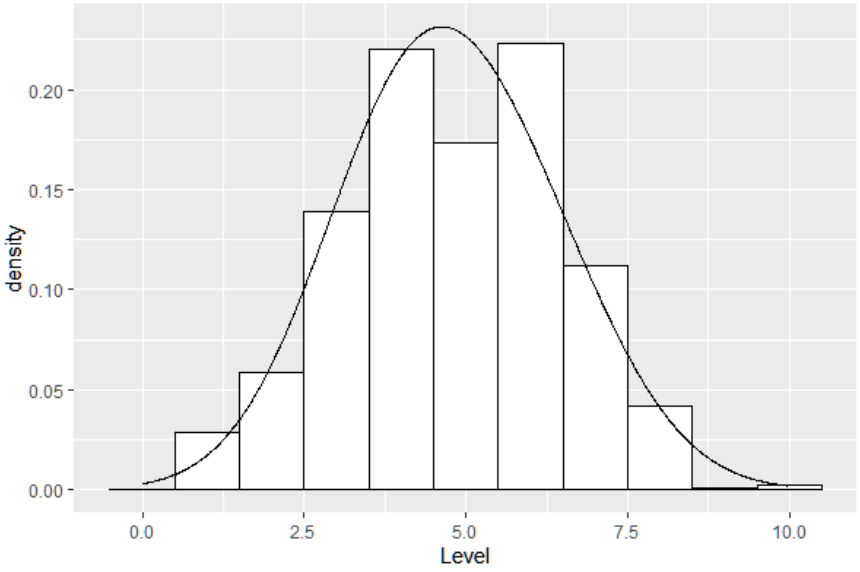


WAIC比較

	WAIC	SE	95%信頼区間
収入	4787.8	61.3	[4670.544 , 4910.056]
準拠集団平均 $\delta_l = 3, \delta_u = 1$	4972.5	58.7	[4857.448, 5087.552]
収入ー準拠集団平均 $\delta_l = 3, \delta_u = 1$	4698.9	62.4	[4576.596, 4821.204]

WAICは汎化損失G_nの漸近平均一致推定量

$$
\begin{aligned}
G_n &= -\mathbb{E}_{q(X)}[\log p^*(X)] \\
&= -\mathbb{E}_{q(X)}[\log \mathbb{E}_{p(\theta|x^n)}[p(X|\theta)]] \\
&= -\int q(x) \log \left(\int p(x|\theta) p(\theta|x^n) d\theta \right) dx
\end{aligned}
$$



自由エネルギー比較

- 下方比較モデル: FE = 12190.59
- 収入回帰モデル: FE = 2412.114
- 近傍モデル（l; u 固定）: FE = 2403.903
- 近傍モデル（l 固定, u 推定）: FE = 2404.711
- 近傍モデル（l; u 推定）: FE 収束せず

ベイズ自由エネルギー F_n は周辺尤度 Z_n の対数符号反転

（最尤推定での最大対数尤度のようなもの）

$$F_n = -\log Z_n$$
$$= -\log \int p(x^n|\theta)\varphi(\theta)d\theta$$

まとめ

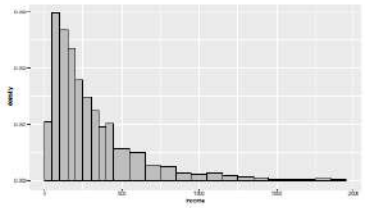
- 客観収入と主観的収入評価のズレは準拠集団モデルで説明できる
- 準拠集団モデルは限界効用逓減則と整合的
- データへのフィットは相対的によい

課題

- 職業情報・空間情報の参照
- 集団別にδを推定せず，分布制約で階層化
- インプリケーションの検証
- 差以外の評価関数形の比較

モデル評価

現実



日本の所得分布 $q(x)$

母集団

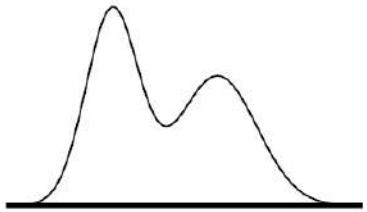
ランダム・サンプリング

$x_1 \quad x_2 \quad \cdots \quad x_n$

実現値（データ）

推定値 $\hat{\theta} = f(x_1, x_2, \ldots, x_n)$

モデルの世界



未知の分布 $q(x)$

ズレの確認

汎化損失

i.i.d.

$X_1 \quad X_2 \quad \cdots \quad X_n$

確率変数

推定量 $\hat{\Theta} = f(X_1, X_2, \ldots, X_n)$

確率モデル $p(x|\theta) \neq q(x)$

予測分布

$p(x|\hat{\theta})$ や $\int p(x|\theta)\underbrace{p(\theta|x^n)}_{\text{posterior}}d\theta$

母集団の推測