



Pythonで体験するベイズ推論

4.3.2~4.4

秋山研究室 M1 伊井 良太



大数の法則

- Z_1, Z_2, \dots, Z_N を、ある確率分布からサンプリングした N 個の独立したサンプルとする

$$\frac{1}{N} \sum_{i=1}^N Z_i \rightarrow E[Z], \quad N \rightarrow \infty$$

- 同じ分布から得られた確率変数の集合の平均は、その分布の期待値に収束する

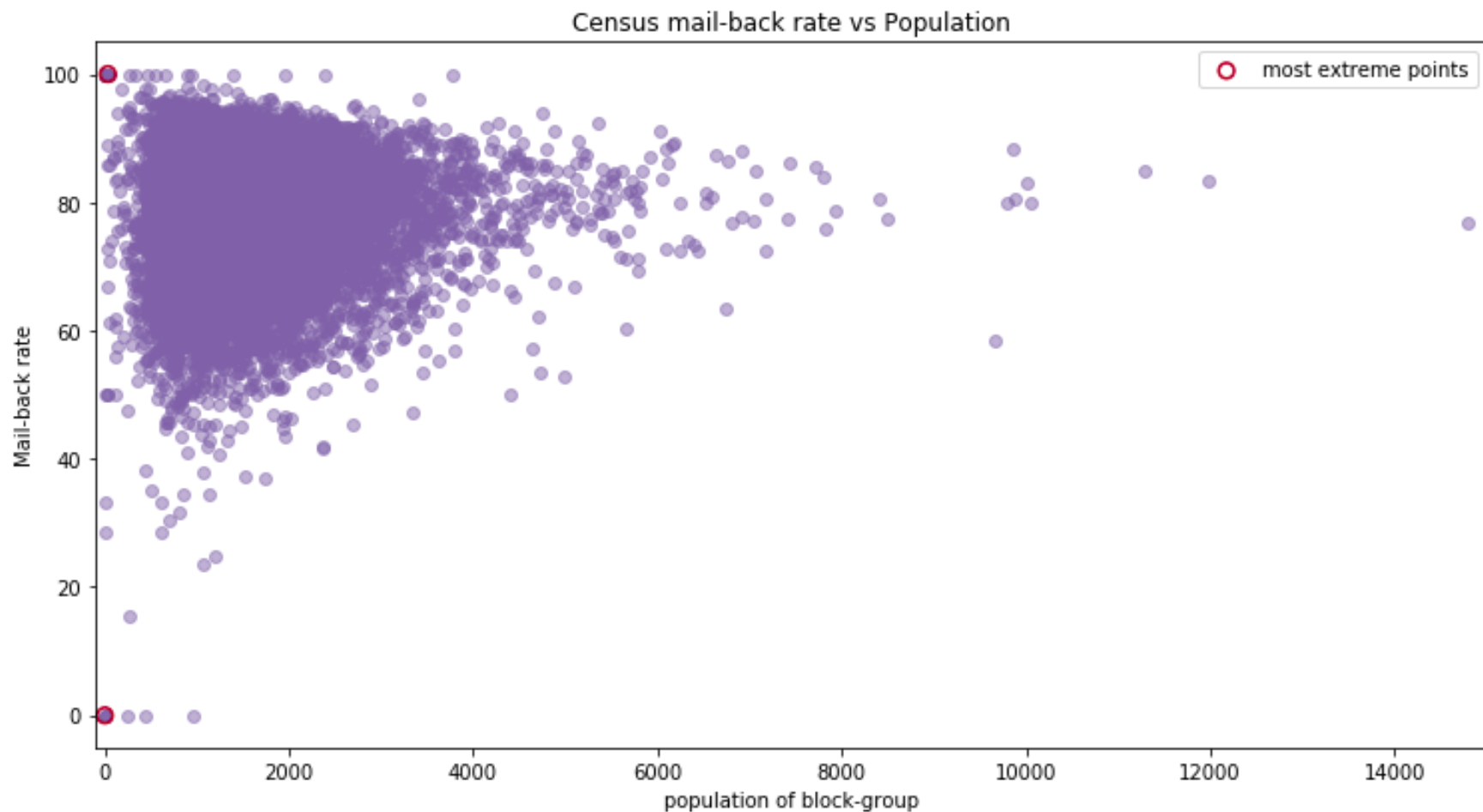
Kaggleのデータセット

- 2010年のアメリカ国勢調査で、町のブロックに相当する調査の単位で集計
- 国勢調査の項目から国勢調査返送率をブロック単位で予測

Mail_Return_Rate_CEN_2010	Tot_Population_CEN_2010
74.5	686
85.799999999999997	1112
77.099999999999994	1409
90.0	1892
77.900000000000006	1134
85.400000000000006	915
84.900000000000006	1181
78.0	1668
80.0	5791
85.799999999999997	



人口に対する国勢調査返送率



アイテムのソート

- 評価者が一人しかいない商品の信頼度

カスタマーレビュー

★★★★★ 1

5つ星のうち5.0

星5つ 100%

星4つ 0%

星3つ 0%

星2つ 0%

星1つ 0%

他のお客様にも意見を伝えましょう

カスタマーレビューを書く

[すべてのカスタマーレビューを見る\(1\)](#)

2人なら？3人なら？



商品の本当の価値を
反映していない

- 信頼できないコメントが上位にきて、価値のあるコメントは「次のページへ」をクリックしないと見つからない

Reddit

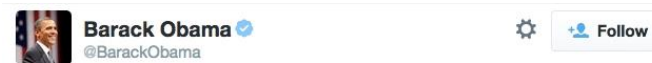


- アメリカの掲示板サイトで、ニュース記事、画像のリンクやテキストを投稿し、コメントをつけることが可能
- 各コメントにupvoteとdownvoteができる
- Redditはデフォでコメントを降順にソート
- どのように並べるべきか

↑ [-] **dvsbastard** 421 points 3 hours ago
↓ I like the idea... But also: <http://i.imgur.com/AqDCn.png>
permalink report reply

↑ [-] **Mrow** 89 points 2 hours ago
↓ Thanks for clearing that up, man.
permalink parent report reply

↑ [-] **SpudgeBoy** 57 points 2 hours ago
↓ This is a good resolution to the problem.
permalink parent report reply



President Obama is answering your questions in a #Reddit AMA, starting right now: [OFA.B0/nNoMPG](https://www.youtube.com/watch?v=OFA.B0/nNoMPG),





どのコメントがベストか

- 人気度...Count(upvote)
 - downvoteがその倍あったらベストかどうか怪しい
- 投票の差...upvote-downvote
 - upvoteがたくさんあるであろう最も古いコメントがトップになりやすい
- 時間調整...difference/the age of the submission
 - 100秒後 99 upvotes < 1秒後 1 upvote
- 比率...upvote/(upvote+downvote)
 - 999 upvotes and 1 downvote < 1 upvote


True Upvote Ratio


- *true upvote ratio* \neq *observed upvote ratio*
- 「そのコメントに誰かがupvoteを投票する潜在的な確率」
- upvote比率の事前分布を決めたい
 - データの歪み...極端な比率をもつコメントが大半
 - データの偏り...可愛い動物の写真と政治の議論とのコメント傾向の違い






一様分布を事前分布に使用する


Reddit Comment Example

 **reddit** ANNOUNCEMENTS 個のコメント


 This is an archived post. You won't be able to vote or comment.


0





reddit changes: individual up/down vote counts no longer visible, "% like it" closer to reality, major improven
Deimorz [A] が 3年前 * 投稿  x3

"Who would downvote this?" It's a common comment on reddit, and is fairly often followed up by someone explaining that reddit "fuzzes" the votes on everything by adding fake votes to posts in order to make it more difficult for bots to determine if their votes are having any effect or not. While it's always been a necessary part of our anti-cheating measures, there have also been a lot of negative effects of making the specific up/down counts visible, so we've decided to remove them from public view.



CamDavidsonPilon commented on Jul 4 2014 Owner

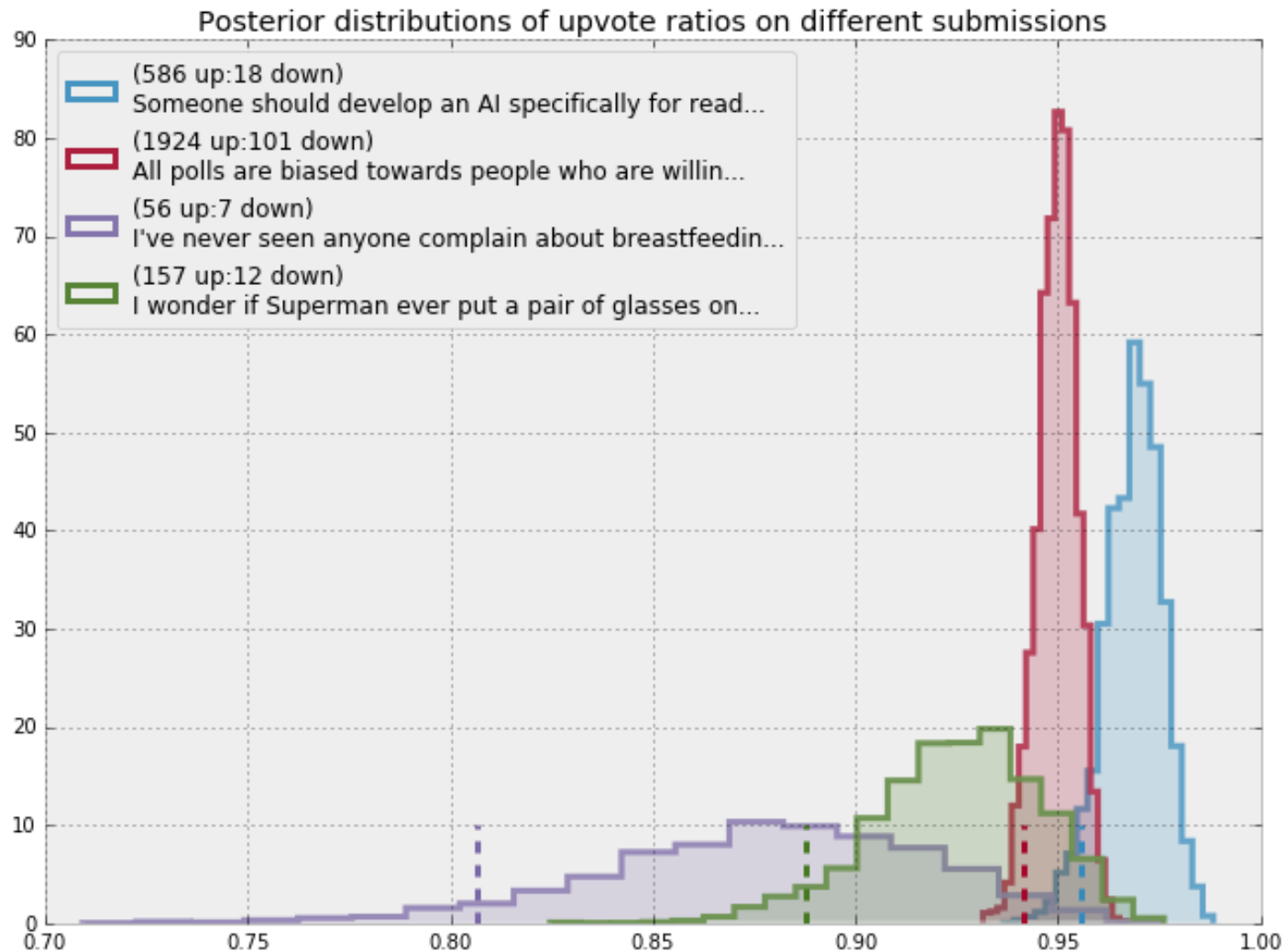
Good discussion, and very sad to see the scores go. I'll keep the example, but add a large Notice describing the current state. I'll probably also disallow the code to be executable, and just translate it to markdown.



コメントのソート

- 95%信用下限(95% least plausible value)
 - 95%信頼区間の下限值
- upvote比率が高くなりそうなコメント
 - プロット中で95%信用下限が1に近いコメント
- upvote比率が同じ二つのコメントが与えられた場合、より多くの投票をもつコメントのほうを良いとみなす
- 投票数が同じ場合、upvoteが多いほうを良いとみなす

upvote比率の事後分布と95%信用下限



ベータ分布

- 確率密度関数

$$f(x; a, b) = \begin{cases} \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1} & (0 < x < 1) \\ 0 & (\text{otherwise}) \end{cases}$$

- ベータ関数 B は正規化定数であり、次式で書ける

$$B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx$$

- 期待値と分散

$$E[X] = \frac{a}{a+b} \quad V[X] = \frac{ab}{(a+b)^2 (a+b+1)}$$

ベータ分布の平均と分散

ベータ関数の積分公式：
$$\int_0^1 x^\alpha (1-x)^\beta = \frac{\alpha! \beta!}{(\alpha + \beta + 1)!}$$

共役事前分布

$$f(\theta | x) \propto f(x | \theta) f(\theta)$$

尤度に二項分布、事前分布にベータ分布を入れる

$$\binom{n}{x} \theta^x (1 - \theta)^{n-x} \times B(a, b)^{-1} \theta^{a-1} (1 - \theta)^{b-1}$$

$$\propto \theta^{x+a-1} (1 - \theta)^{n-x+b-1}$$

$$= \theta^{a'-1} (1 - \theta)^{b'-1}$$

$$a' = x + a$$

$$b' = n - x + b$$

事後分布はパラメータ (a', b') の
ベータ分布になる

信用下限の近似下界

- 事前分布にベータ分布(パラメータ $a = 1, b = 1$)
- 尤度に観測 $u, N = u + d$ の二項分布



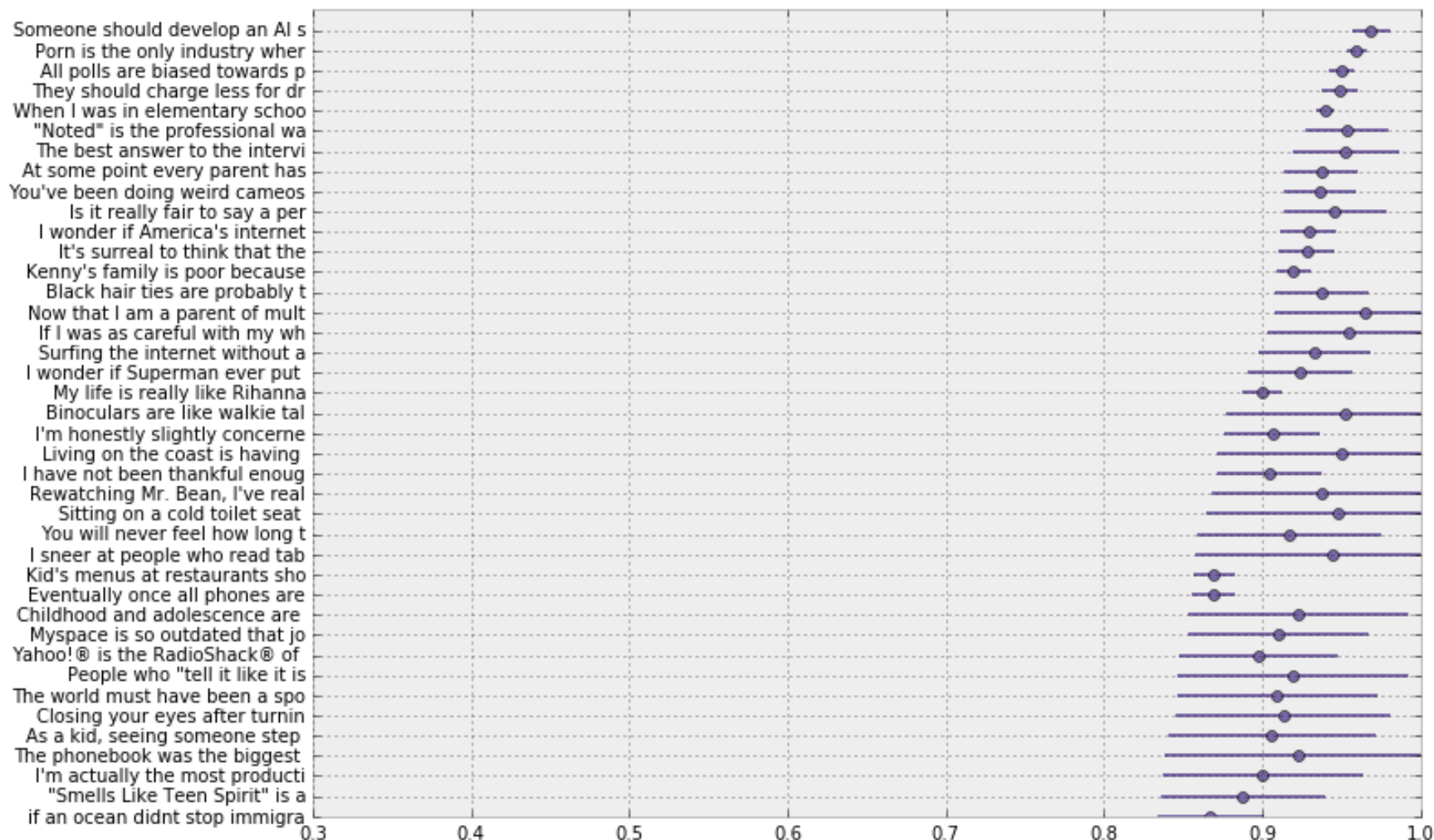
- 事後分布はパラメータ $a' = 1 + u, b' = 1 + (N - u) = 1 + d$ のベータ分布
- 確率が0.05よりも小さくなるような x の値を求める

$$0.05 = \Phi\left(\frac{(x - \mu)}{\sigma}\right) \quad \Phi \text{は正規分布の累積分布}$$

$$\frac{a}{a+b} - 1.65 \sqrt{\frac{ab}{(a+b)^2(a+b+1)}} \quad \begin{array}{l} a \text{はupvoteの数、} \\ d \text{はdownvoteの数} \end{array}$$



下界でソートした場合の上位のコメント



評価システムへの拡張

- 五つ星をつけるような場合は？
- *upvote/downvote*は0/1の2値問題
- N つ星評価システムはその連続版で、
 n つ星がついたら n/N がついたとみなす

$$\frac{a}{a+b} - 1.65 \sqrt{\frac{ab}{(a+b)^2(a+b+1)}}$$

ここで

$$a = 1 + S$$

$$b = 1 + N - S$$

N は評価したユーザーの数、 S はすべての評価の和

演習問題 1

- $X \sim \text{Exp}(4)$ のとき、
 $E[\cos X]$ と $E[\cos X | X < 1]$ を推定せよ。
- 必要であれば、以下のモジュールをインポートしてよい。
 - *import scipy.stats*
 - *from numpy import cos*

演習問題 2 - 1

- 以下の表は、アメリカンフットボールのフィールド・ゴール・キッカーをキックの成功率でランク付けしたものである。この表の誤りを述べよ。

Rank	Kicker	Make %	Number of Kicks
1	Garrett Hartley	87.7	57
2	Matt Stover	86.8	335
3	Robbie Gould	86.2	224
4	Rob Bironas	86.1	223
5	Shayne Graham	85.4	254
...	
51	Dave Rayner	72.2	90
52	Nick Novak	71.9	64
53	Tim Seder	71.0	62
54	Jose Cortez	70.7	75
55	Wade Richey	66.1	56

演習問題 2 - 2

- 右の表は、
使用している
プログラミング言語
によるプログラマーの
平均年収の違いである。
上位と下位を見て、
何がわかるかを述べよ。

Language	Average Household Income (\$)	Data Points
Puppet	87,589.29	112
Haskell	89,973.82	191
PHP	94,031.19	978
CoffeeScript	94,890.80	435
VimL	94,967.11	532
Shell	96,930.54	979
Lua	96,930.69	101
Erlang	97,306.55	168
Clojure	97,500.00	269
Python	97,578.87	2314
JavaScript	97,598.75	3443
Emacs Lisp	97,774.65	355
C#	97,823.31	665
Ruby	98,238.74	3242
C++	99,147.93	845
CSS	99,881.40	527
Perl	100,295.45	990
C	100,766.51	2120
Go	101,158.01	231
Scala	101,460.91	243
ColdFusion	101,536.70	109
Objective-C	101,801.60	562
Groovy	102,650.86	116
Java	103,179.39	1402
XSLT	106,199.19	123
ActionScript	108,119.47	113