

四捨五入したパーセンテージが何個あれば
共通する分母を逆算して求まる値に
確信が高く持てるかについての
ベイズ推定の考え方による考察

2022-11-07 (月) 下野寿之

ベイズの式(一般論)

$$\text{Prob}(h_i|e) \propto \text{Prob}(h_i) \times \text{Prob}(e|h_i)$$

この数式の証明は容易。下記の応用のための理解が重要。

- $\text{Prob}(h_i)$ の項は「**事前確率**」と呼ばれ、異なる仮説間の比が本質的である。
(自然で尤もな確率を、主観的でも、当てはめる必要がある。)
- $\text{Prob}(e|h_i)$ は「**尤度**」と呼ばれ、通常は一意に数理的に決定可能な方法で計算する。
- 証拠 e に対して、上記2個が定まればベイズの式で異なる仮説 $h_i (i=1,2,3\dots)$ がそれぞれどれだけ可能であるかが左辺の「**事後確率**」 $\text{Prob}(h_i|e)$ である。
- 仮説 h_i で異なるものが同時に起きない場合は、事前確率も事後確率も総和は100%であり、この性質を計算でよく利用する。
i.e. 算出された比(もしくは連比)で、100%を**分配**する。

考察の方法 (概念的な流れ) :

1. 分母逆算器により、割合を四捨五入した値M個から、分母候補 S を算出する仕組みを用意する。
2. 事前確率 P_{in} を可能な分母の全体の上に決める。
分母候補 各値の事後確率 P_{out} が算出可能になる。
3. 分母Dを確率変数として $P_{in}(D)$ の確率で生成し、
分子 $1 \leq N_i \leq D-1$ を M 個生成し(離散一様分布で i.i.d.)、
各 N_i/D の四捨五入値(0.01 の位まで)全体から S を得る。
4. この D を含む S の上での P_{out} が定義できるので、
元の D での値 $P_{out}(D)$ を**擬似正解確率** y とする。
その y の値の分布を、異なる M=1,2,3... で考察。

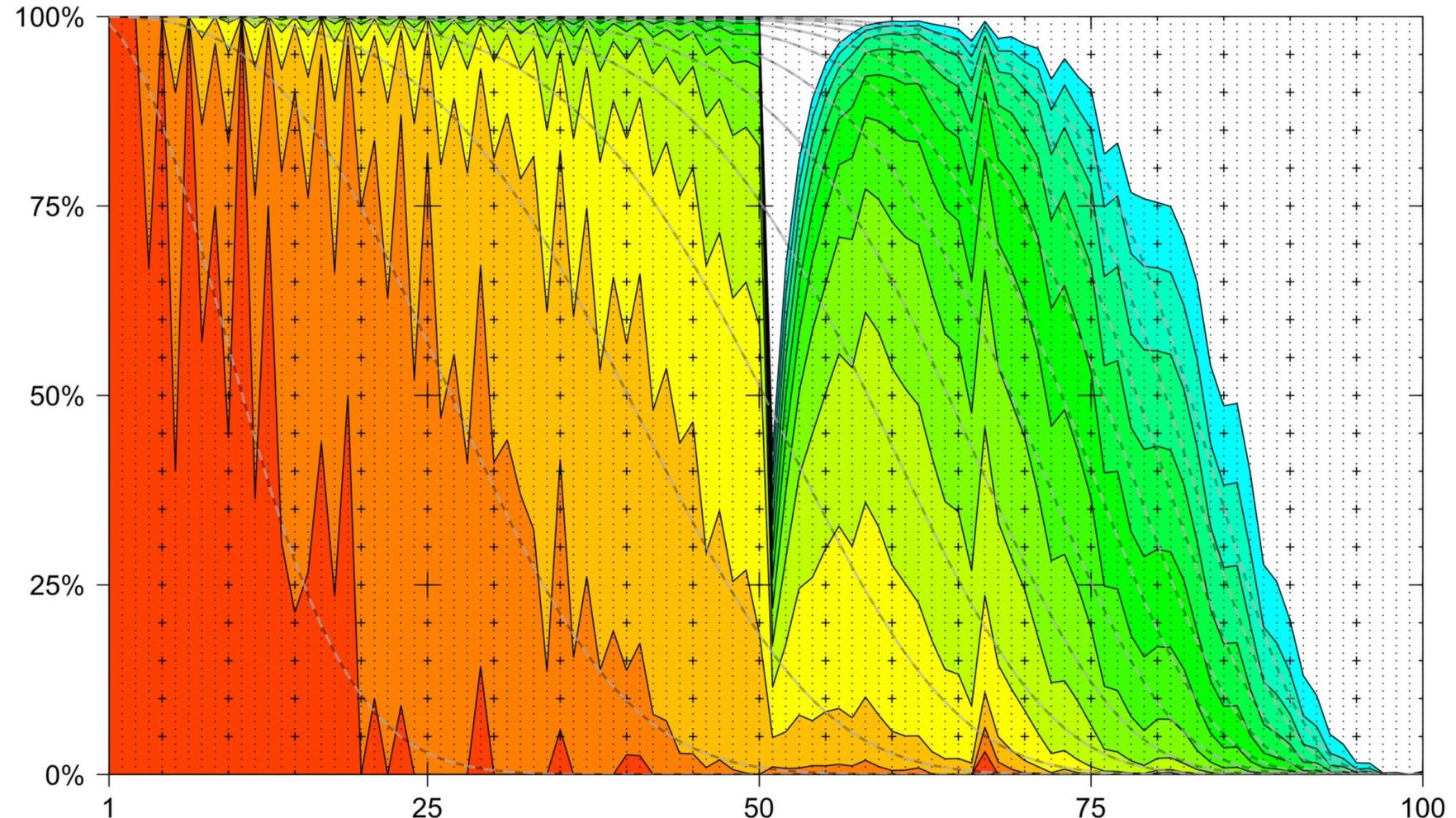
最初に設ける具体的な仮定：

1. 分子を分母で割って求めた割合の値 R_i を、四捨五入した結果 F_i は、小数点以下2桁とする。
2. 分母 D は **2から50以下** の整数とする。
 - 51以上を考えない理由 (※下記2点は厳密な議論ではない):
 - 50以下の限定で、 $M \geq 7$ なら、分母決定確率 $\geq 97\%$ (後述)。
 - 51, 52..を含めると49, 48..と高確率で識別困難ゆえ別議論を要す。
 - 分母1の場合を考える必要もないと考えられる。
3. 事前分布 P_{in} は 分母の値の逆数(1/D) に比例させる。
4. 分子の分布は **1からD-1** の整数とする。
 - 分母 D に対し、分子の値が0と D の場合は考察不要。
 - 尤度は $1/(D-1)^M$ に比例することになる。
 - もしも分母の値が D なら、各 F_i の値を引き起こす確率は $1/(D-1)$ 。

ここまで考察方法の採用理由：

1. まず、**共通分母の逆算は経験的に容易であった**:
 - 特に割合近似値が3~5個以上ある場合。
 - 分母候補が複数あっても**最小値が元の分母の値**のことが、なぜが多い。
 2. しかし、分母の値としてDが候補なら**Dの倍数も全て候補**である。
 - M個の近似値から逆算した場合に、分母も分子もn倍した場合も候補になるが、事象「分子全てがnの倍数」は n^M 倍の確率を考えると、自然だし都合が良い。
 - このことは、**尤度**の考え方で、正当化できる。
 3. 分母Dが2,3,4,..になる確率と20~29, 30~39, 40~49..になる確率をそれぞれ、大体揃えたい。
 - 分母の値が各整数で等確率だと、桁数の多い数ばかりとなり不自然。
 - 分母の値について、その出現確率はその逆数に比例すると仮定して、**事前分布**(事前確率の割り当て)を与えると、計算上も都合が良い。
 - 分母の値の探索範囲として、51以上を含めると不具合(後述)があった。
 4. 分子の値の分布の仮定について:
 - 割合が整数の場合は考える必要なし。
 - 分母の値Dに対して、法Dで考えると、1~D-1の整数のみを考えれば良い。
 - 「同様に確からしい」の考えを採用して、一様離散分布を**事前分布**とした。
- 「M個の割合の総和が100%」の設定も考えたいが、今回は考慮外。

別の考察1: 近似値の個数Mがどれくらいあれば元の分母の値が求まる確率が高まるか



横軸は分母D。 $M=1,2,3,\dots,12$ に対して、左下の赤系色から水色に対応。縦軸は、「 M 個の異なる割合近似値をランダムに与えて分母候補を逆算した時に、その候補の最小値がDに一致する確率」。

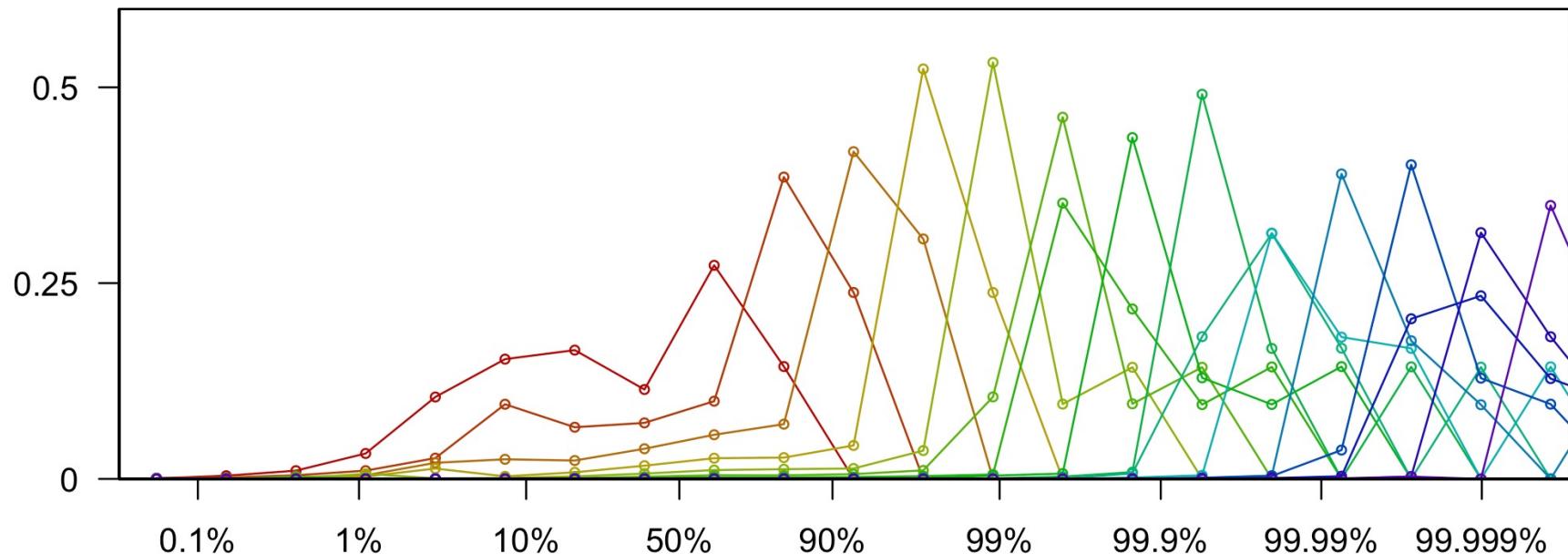
別の考察2: 分母が51の割合の近似値は49の場合と混同しやすい。52は48と同様。53は47と同様。

	.03	.06	.09	.11	.14	.17	.20	.23	.26	.29	.31	.34	.37	.40	.43	.46	.49															
35:	.03	.06	.09	.11	.14	.17	.20	.23	.26	.29	.31	.34	.37	.40	.43	.46	.49															
36:	.03	.06	.08	.11	.14	.17	.19	.22	.25	.28	.31	.33	.36	.39	.42	.44	.47	.50														
37:	.03	.05	.08	.11	.14	.16	.19	.22	.24	.27	.30	.32	.35	.38	.41	.43	.46	.49														
38:	.03	.05	.08	.11	.13	.16	.18	.21	.24	.26	.29	.32	.34	.37	.39	.42	.45	.47	.50													
39:	.03	.05	.08	.10	.13	.15	.18	.21	.23	.26	.28	.31	.33	.36	.38	.41	.44	.46	.49													
40:	.03	.05	.08	.10	.13	.15	.18	.20	.23	.25	.28	.30	.33	.35	.38	.40	.43	.45	.48	.50												
41:	.02	.05	.07	.10	.12	.15	.17	.20	.22	.24	.27	.29	.32	.34	.37	.39	.41	.44	.46	.49												
42:	.02	.05	.07	.10	.12	.14	.17	.19	.21	.24	.26	.29	.31	.33	.36	.38	.40	.43	.45	.48	.50											
43:	.02	.05	.07	.09	.12	.14	.16	.19	.21	.23	.26	.28	.30	.33	.35	.37	.40	.42	.44	.47	.49											
44:	.02	.05	.07	.09	.11	.14	.16	.18	.20	.23	.25	.27	.30	.32	.34	.36	.39	.41	.43	.45	.48	.50										
45:	.02	.04	.07	.09	.11	.13	.16	.18	.20	.22	.24	.27	.29	.31	.33	.36	.38	.40	.42	.44	.47	.49										
46:	.02	.04	.07	.09	.11	.13	.15	.17	.20	.22	.24	.26	.28	.30	.33	.35	.37	.39	.41	.43	.46	.48	.50									
47:	.02	.04	.06	.09	.11	.13	.15	.17	.19	.21	.23	.26	.28	.30	.32	.34	.36	.38	.40	.43	.45	.47	.49									
48:	.02	.04	.06	.08	.10	.13	.15	.17	.19	.21	.23	.25	.27	.29	.31	.33	.35	.38	.40	.42	.44	.46	.48	.50								
49:	.02	.04	.06	.08	.10	.12	.14	.16	.18	.20	.22	.24	.27	.29	.31	.33	.35	.37	.39	.41	.43	.45	.47	.49								
50:	.02	.04	.06	.08	.10	.12	.14	.16	.18	.20	.22	.24	.26	.28	.30	.32	.34	.36	.38	.40	.42	.44	.46	.48	.50							
51:	.02	.04	.06	.08	.10	.12	.14	.16	.18	.20	.22	.24	.25	.27	.29	.31	.33	.35	.37	.39	.41	.43	.45	.47	.49							
52:	.02	.04	.06	.08	.10	.12	.13	.15	.17	.19	.21	.23	.25	.27	.29	.31	.33	.35	.37	.38	.40	.42	.44	.46	.48	.50						
53:	.02	.04	.06	.08	.09	.11	.13	.15	.17	.19	.21	.23	.25	.26	.28	.30	.32	.34	.36	.38	.40	.42	.43	.45	.47	.49						
54:	.02	.04	.06	.07	.09	.11	.13	.15	.17	.19	.20	.22	.24	.26	.28	.30	.31	.33	.35	.37	.39	.41	.43	.45	.47	.49						
55:	.02	.04	.05	.07	.09	.11	.13	.15	.16	.18	.20	.22	.24	.25	.27	.29	.31	.33	.35	.36	.38	.40	.42	.44	.45	.47	.49					
56:	.02	.04	.05	.07	.09	.11	.13	.14	.16	.18	.20	.21	.23	.25	.27	.29	.30	.32	.34	.36	.38	.39	.41	.43	.45	.46	.48	.50				
57:	.02	.04	.05	.07	.09	.11	.12	.14	.16	.18	.19	.21	.23	.25	.26	.28	.30	.32	.33	.35	.37	.39	.40	.42	.44	.46	.47	.49				
58:	.02	.03	.05	.07	.09	.10	.12	.14	.16	.17	.19	.21	.22	.24	.26	.28	.29	.31	.33	.34	.36	.38	.40	.41	.43	.45	.47	.48	.50			
59:	.02	.03	.05	.07	.08	.10	.12	.14	.15	.17	.19	.20	.22	.24	.25	.27	.29	.31	.32	.34	.36	.37	.39	.41	.42	.44	.46	.47	.49			
60:	.02	.03	.05	.07	.08	.10	.12	.13	.15	.17	.18	.20	.22	.23	.25	.27	.28	.30	.32	.33	.35	.37	.38	.40	.42	.43	.45	.47	.48	.50		
61:	.02	.03	.05	.07	.08	.10	.11	.13	.15	.16	.18	.20	.21	.23	.25	.26	.28	.30	.31	.33	.34	.36	.38	.39	.41	.43	.44	.46	.48	.49		
62:	.02	.03	.05	.06	.08	.10	.11	.13	.15	.16	.18	.19	.21	.23	.24	.26	.27	.29	.31	.32	.34	.35	.37	.38	.39	.40	.42	.44	.45	.47	.48	.50
63:	.02	.03	.05	.06	.08	.10	.11	.13	.14	.16	.17	.19	.21	.22	.24	.25	.27	.29	.30	.32	.33	.35	.37	.38	.40	.41	.43	.44	.46	.48	.49	
64:	.02	.03	.05	.06	.08	.09	.11	.13	.14	.16	.17	.19	.20	.22	.23	.25	.27	.28	.30	.31	.33	.34	.36	.38	.39	.41	.42	.44	.45	.47	.48	.50
65:	.02	.03	.05	.06	.08	.09	.11	.12	.14	.15	.17	.18	.20	.22	.23	.25	.26	.28	.29	.31	.32	.34	.35	.37	.38	.40	.42	.43	.45	.46	.48	.49

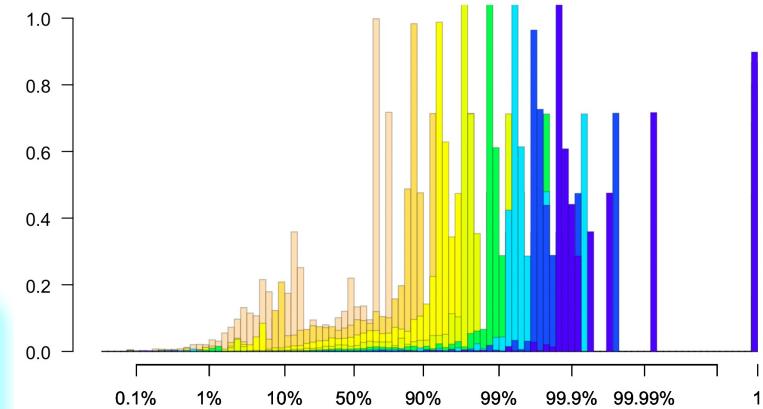
左端の数は分母である。その右に、整数の分子に対する割合の近似値として、小数点とそれ以下の2桁を並べた。

中央の行の50から注目すると、上に行くと数の並びの個数は減少していく様子が、下に向かって数の並びの個数が増加して割り込んで入っていく様子が、かなり対称的に並んでいる。

割合近似値の個数 M が $1, 2, \dots, 16$ の各場合の擬似正解確率 γ のヒストグラム

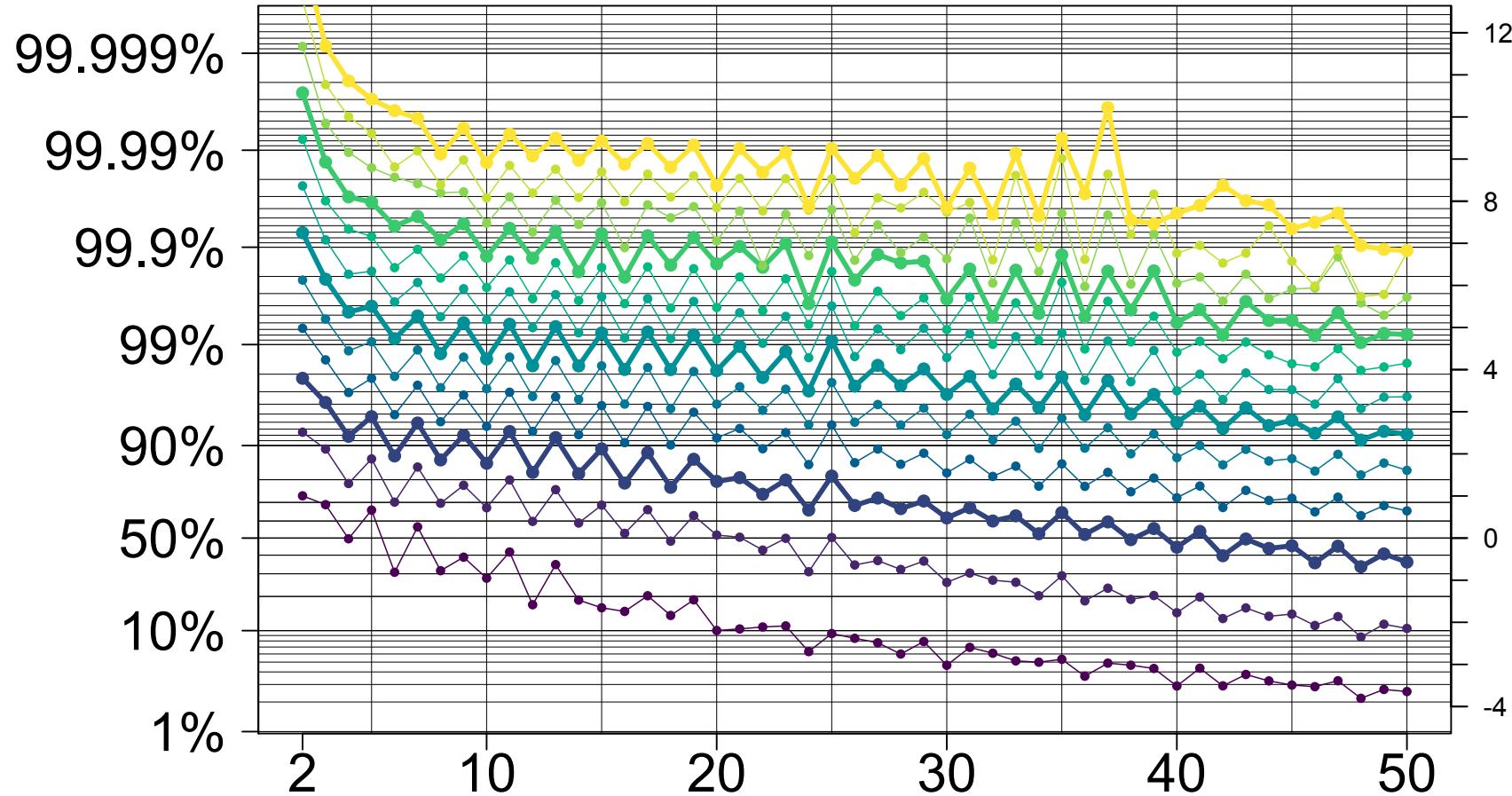


- 各 M に対し **100万回のモンテカルロ法** で γ の値を計算。
- 横軸は γ の値を表示であるが、**描画位置はロジット変換**。
- 左の赤系色から、右の青系色にかけて、 M が 1 ずつ増加。
- 16 個のヒストグラムに対して：
 - 視認性の都合で折れ線グラフに置換。
 - ロジット値で整数で区切った。
- 擬似正解確率** である γ の値は **ほとんどの場合**：
 - M が 1 増加するにつれ、ロジット値は約 1 増加する。
 - 割合近似値の個数 M が $1, 2, 3$ の場合だと、99% 未満。
 - $M \geq 7$ だと 99% 以上。 $M \geq 4$ でも 90% 以上。**



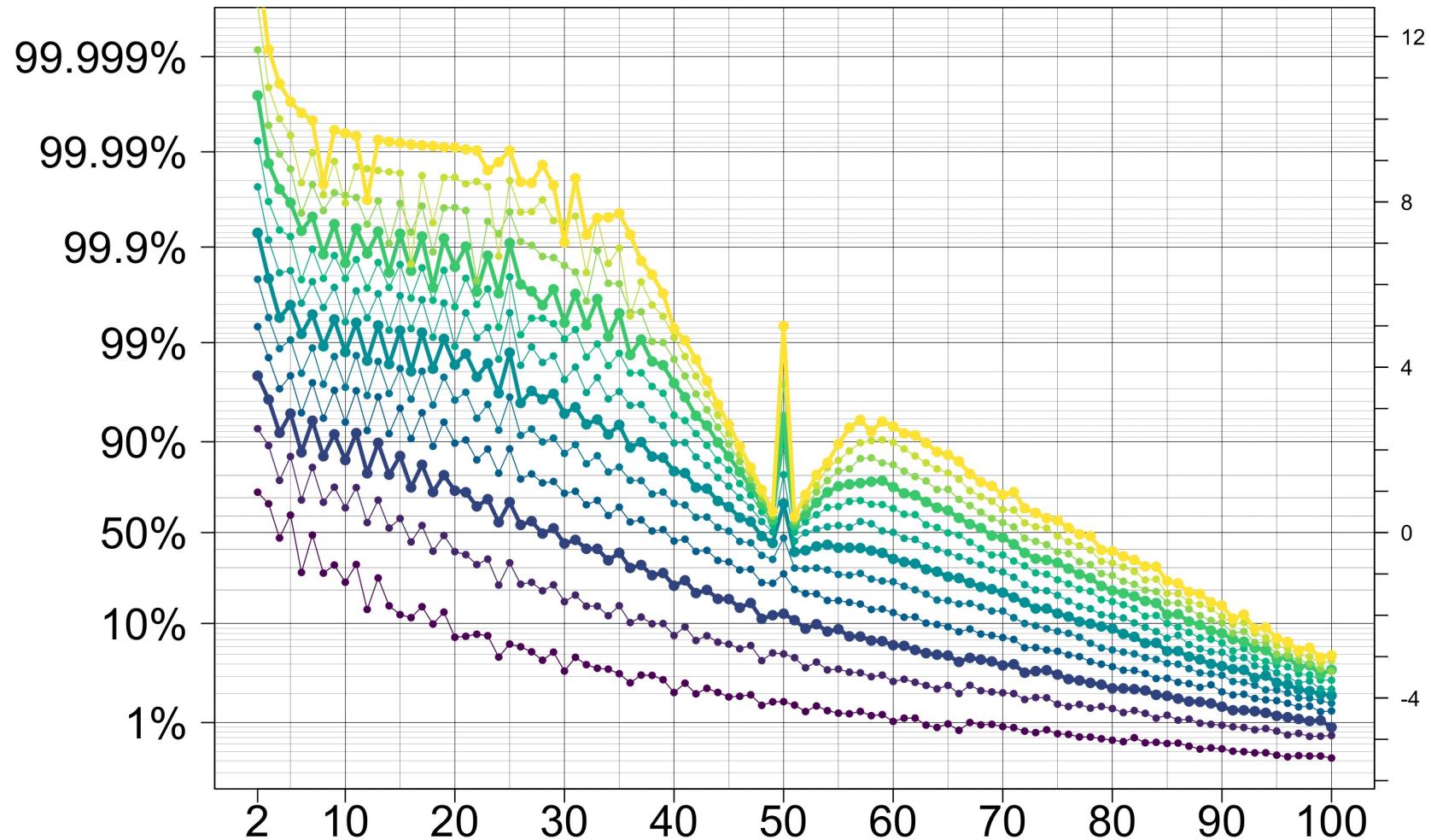
▲ $M=1, \dots, 8$ に対してヒストグラムを異なる色で描いて重ねた場合。
横方向はロジット値で 0.2 刻み。

元の分母の値ごとの疑似正解率の平均値



- $M=1,\dots,12$ について下(青紫)から上(黄)に、それぞれの折れ線グラフを描画した。
- 横軸は元の分母 D_0 。縦軸は、 D_0 の値ごとに疑似正解率の平均値(左)とそのロジット値(右)。
- 有限試行に伴う誤差も一部見える: M の値ごとに丁度100万回、さらに各 D_0 ごとに5000回以上は試行している。計算上、数個が0.5以下で残りが1の場合があり、大きな誤差が発生。
- 各折れ線グラフが全体として右下がりなのは、分母が小さいと事前確率が高いため。
- 元の分母 D_0 の偶奇に伴う凸凹が大半の箇所で見える。 D_0 が4以下で急激に低下し、5で $M \leq 7$ で増加。
- $M=4$ であっても $D \geq 40$ で疑似正解率は60%～80%に低下。 $M \geq 6$ でどの D_0 でも90%以上を確保可能。 9

参考: 分母を50以下の限定にせず
2~100までにした場合



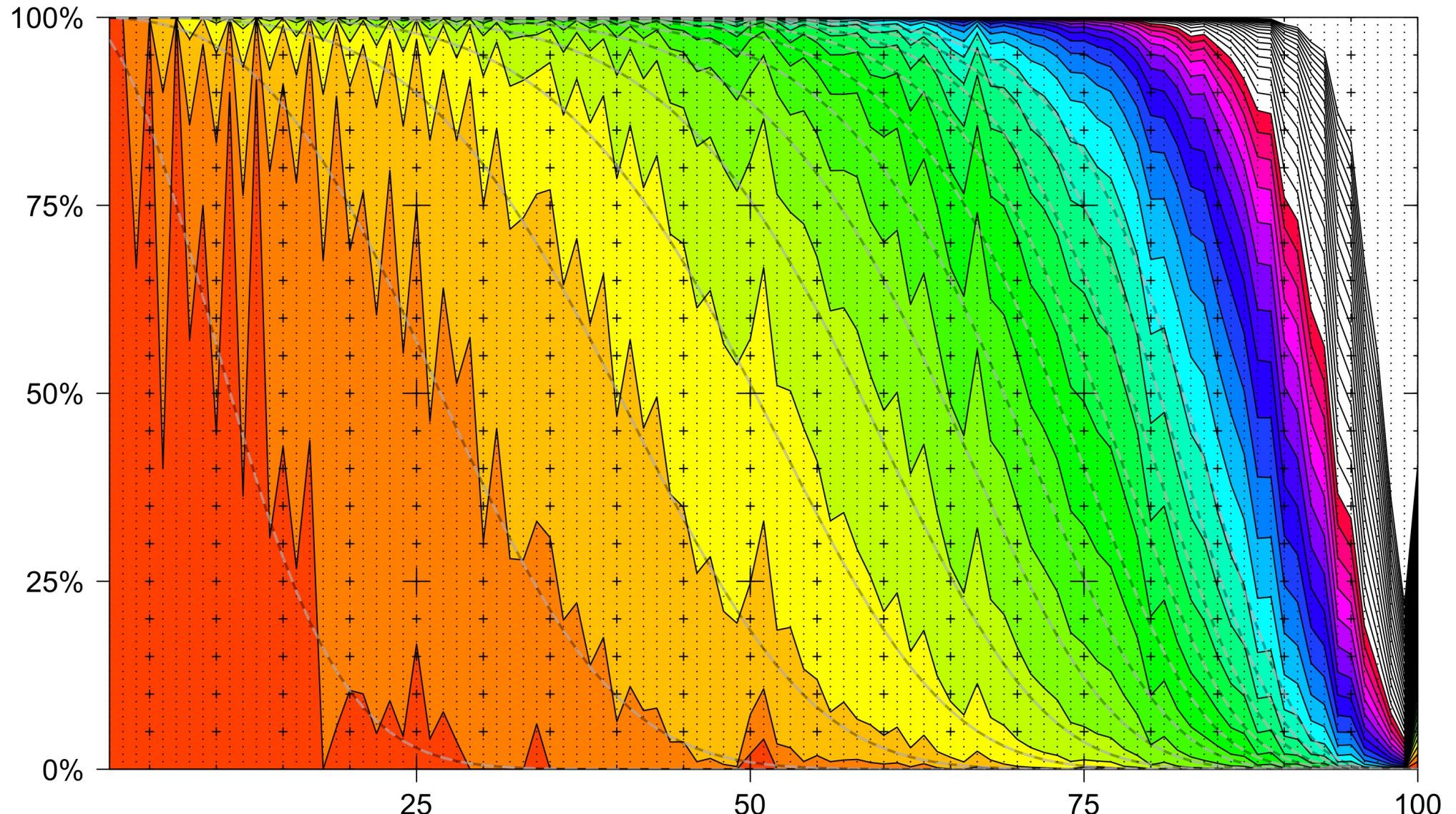
- 事前分布は分母に反比例のままにした。各Mについて20万回ランダムに試行。
- 50以下に限定した場合に比べ、M=4でも6でも、分母を決めにくくなる。
- 従って、分母を50以下に限定することは、大きな意味があったと考えられる。 10

まとめ

- 整数に四捨五入したパーセンテージがM個あれば、それらに共通する**50以下の分母を決める問題**を考えると、
 - $M \leq 3$ なら、分母を確実に決めることはあまり出来ない。
 - $M \geq 4$ なら90% (分母>20なら60%)以上の確信で求めることが出来る。
- 上記を得るためにやや恣意的な仮定を与えた。
 - 与えるパーセンテージ(割合近似値)にランダム性を仮定。
 - ベイズ統計学の事前分布を、恣意的だがひとつ仮定した。
 - 割合近似値を得る状況に、不自然な仮定を与えないためである。
- 結論を利用するときの注意：
 - ある分母Dに対して、M個の分子の総和がDに等しいと仮定される状況には対応していない。
- 分母が50前後で要注意。考えたいこと：
 - 分母を50から100に限定した場合はどうなるか?
 - 分母が2から100で、事後確率の高い数個の候補でどうなるか?
- 未知の数理的な知見がまだありえそうである。

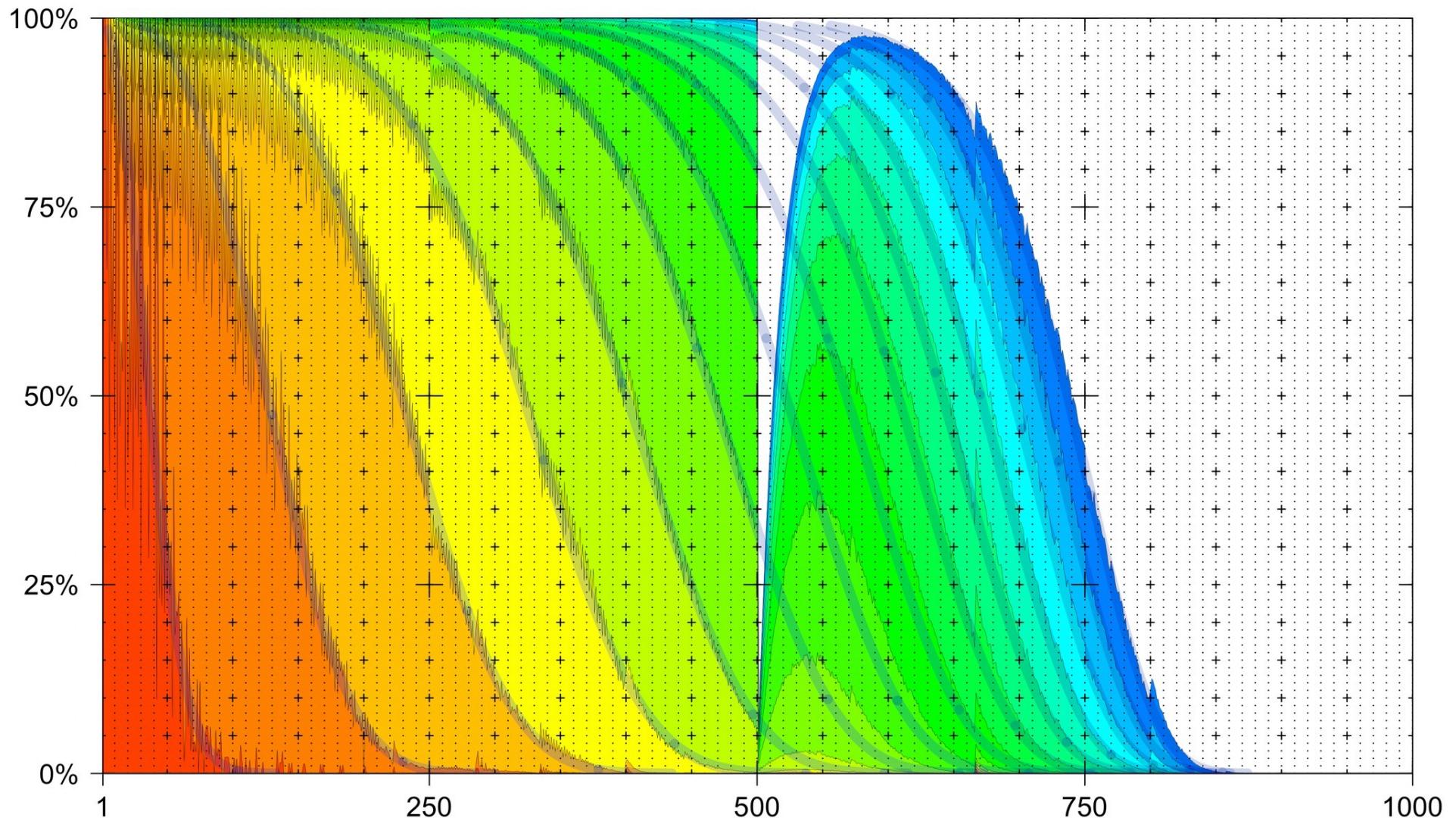
補足

「四捨五入」ではなくて「切り捨て」の場合



四捨五入の場合の、50付近で、50から等距離離れた2個の分母がお互い似たような割合近似値を出力するという現象がなくなることによる効果が現れる。¹³

割合近似値の桁数をさらに1桁増やした場合 (四捨五入して0.1%単位にした場合)



薄い太い線は、 $\prod_{i=1}^{D-1} (1 - (0.001 \times i)^M)$ である。

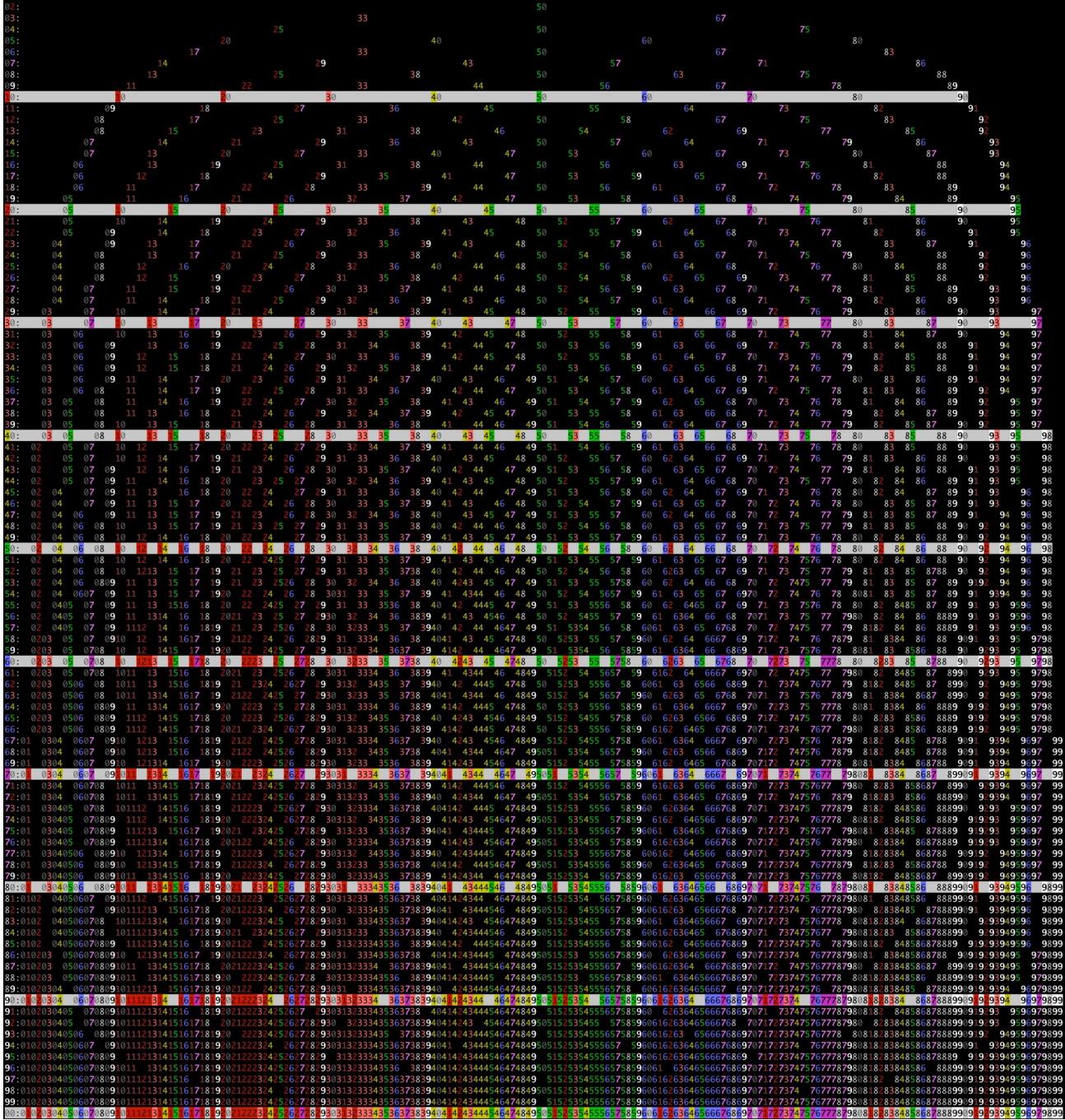
全体の0.1%を占有する各M個の物体が*i*/1000の確率で当たる場合に、
全てが当たることが、どれかの*i*=1,...,D-1で発生する確率である。
14

2から100の分母に対し
割合近似値としてどんな
値が出現するかの、
小数点以下の値を並べた。

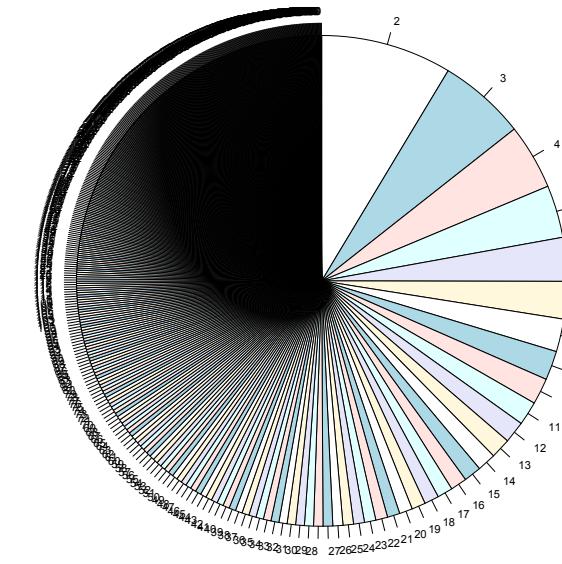
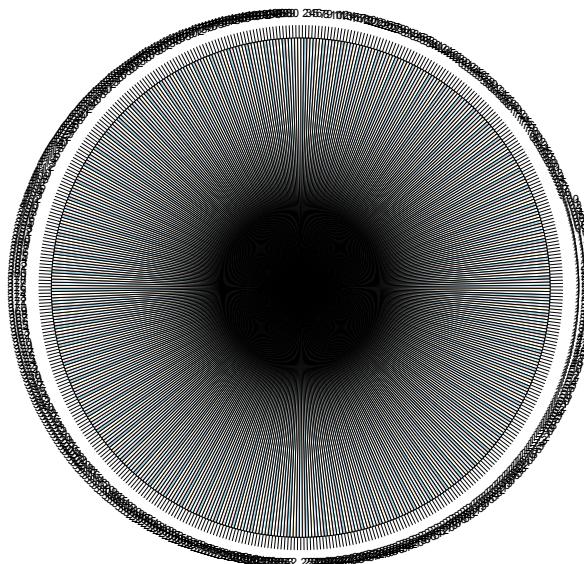
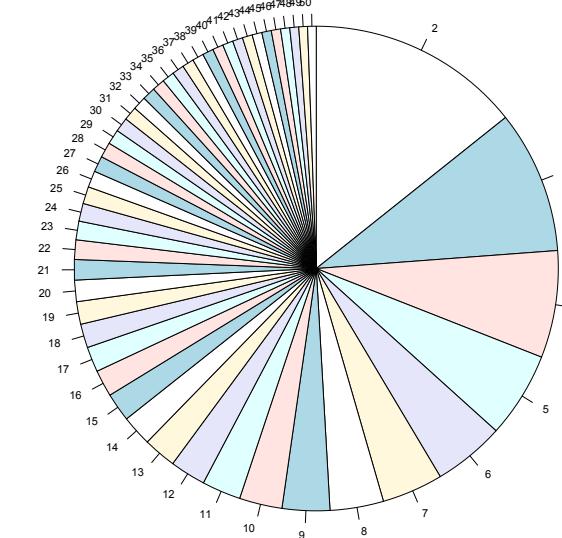
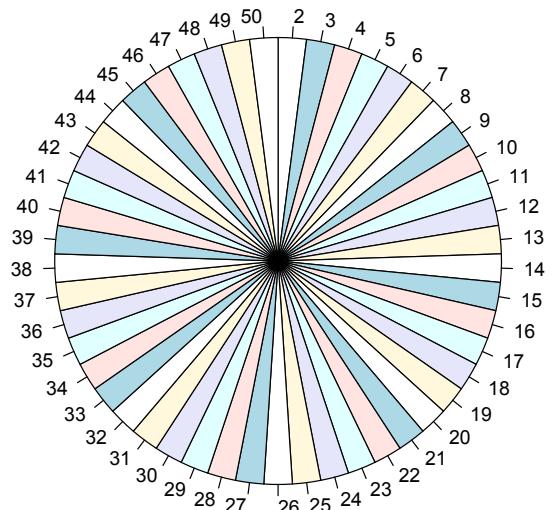
分母Dが1に近い数の場合は、
そういう数の並びは、
空白をD-1箇所でほぼ同じ長
さで切るように並ぶ。

Dが100に近い数の場合は、
数の並びは、穴が100-D個
あって、横にほぼ等間隔だが、
両端の穴は端からその間隔の
半分の所にある。

D=50の横方向の一線を境に、
上下に対称的に50から遠ざかる
につれて、穴が増えたり、
穴が減ったりする箇所が縦方
向によく一致する。



参考: 分母のさまざまな事前分布



考えるべき視点

- この資料の 学術における分野的な位置づけは?
 - 数値解析? (誤差伝搬など)
 - Data Profiling? (データの間違いの発見など)
- 小数点以下3桁(0.1%)単位でやりたかった。
 - 計算時間が1%単位に比べてかなり長くなる。
 - 多くの統計書類は、多くの場合は1%より細かい。
- 考えた数理的な問題を列挙せよ
 - 単体上(四面体,五胞体)で乱数を取ること。
 - 異なりうる形の単体複数を組み合わせた場合(計100%の調査が複数の場合)
 - 出現したカーブが、近似カーブで近似しやすかった。
 - 50前後または500前後についての高めの対称性
- 考えたかった元の問題を時々思い出す
 - ストックしたURLを見てまわると次々とヒントが見つかる←ストックを増やすのは大変。
 - 派生した問題もよく考えたい。しかし元の問題を意識せよ。
- 数値グラフの作図は意外と大変だったことはメモしよう。
 - 一般的な既存のやり方や今まで確立したやり方がすぐ通用しなかった。

さらに行いたいこと

- 1の補数で計算結果を出力すること。
- 分母AとBのそれぞれが生成する割合近似値が
 - どう含まれるかを、 99×99 の行列で示す。
 - その行列の(40～60) × (40～60) 付近を観察せよ。
- 50以下の限定ではなく100以下に緩和すること
 - 当初の問題を解くため。
 - 50以下の限定は、問題の構造を分かりやすく見るための便宜でしかない。

懸案事項

- denomfindで10進文字列で計算している箇所は正しく計算していると長らく思っていたが、そうではないかも。
- Cコンパイラに依存すると perldoc perlnumber に書いてあった。
- 9が20個ある 0.999999999999999999 の例でいろいろ実験したら、確かに私の想定外の動きをする。

他の疑問

- 分母Dに対して、 $1/D$ の事前分布としたが、 $1/(D-1)$ の方が `posteriorp.pl` の計算上は楽かも。