二項分布
binomial distribution

正田 備也

masada@rikkyo.ac.jp

今回のお題

- 簡単な例（コイン投げ）を使って確率を使ってデータをモデル化するとはどういうことかを学ぶ
  - 観測データをモデル化するとはどういうことか
  - ベイズ的にモデル化するとはどういうことか

目次

- ベルヌーイ分布と二項分布
- 最尤推定
- ベイズ的な考え方

ベルヌーイ分布と
二項分布

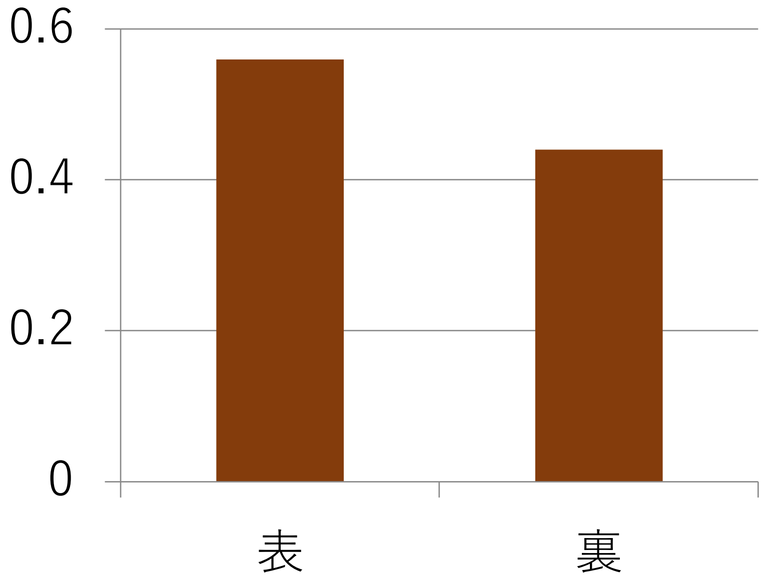
ベルヌーイ分布 Bernoulli distribution

- $V = \{\text{head}, \text{tail}\}$という2種類のitemの集合（例：コイン投げ）
- ベルヌーイ分布はV上の確率分布（離散確率分布）
- ベルヌーイ分布のパラメータ

  ϕ_{head}　：表の出現確率

  ϕ_{tail}　：裏の出現確率

  - $\phi_{\text{head}} + \phi_{\text{tail}} = 1$が成り立つ
    - よって、自由度は1。



二項分布 binomial distribution

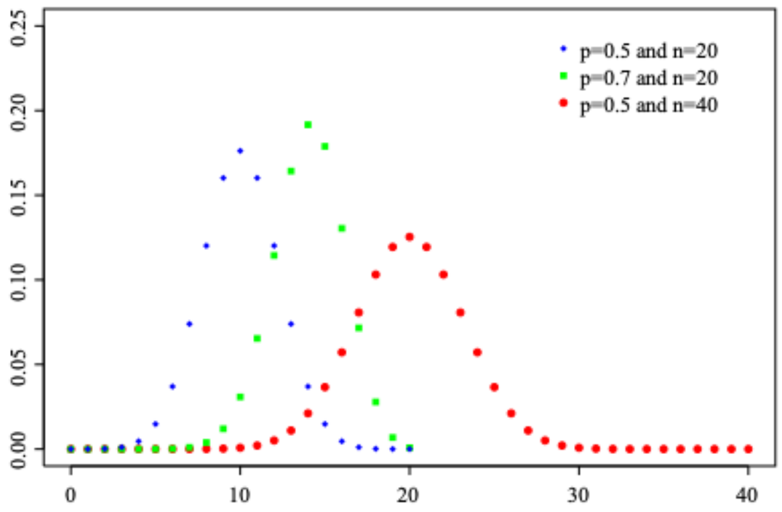
- ベルヌーイ分布は1回のコイン投げのモデリングに使う
- 複数回のコイン投げのモデリングには二項分布を使う
- <u>n回のうち表がk回出る確率は何々</u>と確率を割り振る確率分布
- 二項分布のパラメータ

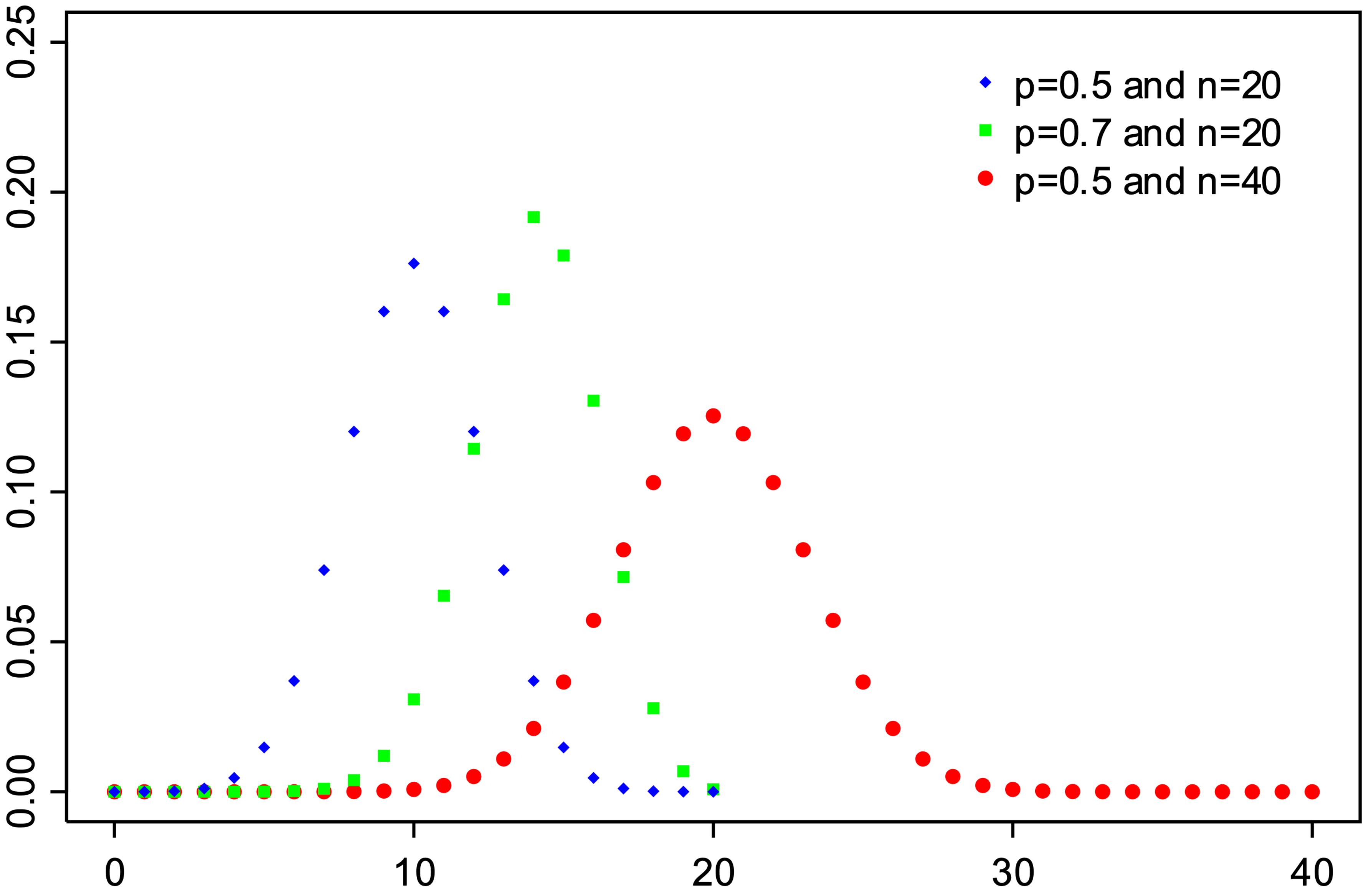
  n ：コインを投げる回数

  ϕ_{head} ：表の出現確率

  ϕ_{tail} ：裏の出現確率

  - $\phi_{\mathrm{head}} + \phi_{\mathrm{tail}} = 1$が成り立つので、自由度は1





https://en.wikipedia.org/wiki/Binomial_distribution#/media/File:Binomial_distribution_pmf.svg

例）5回のコイン投げ (1/3)

- 二項分布は以下の6つの事象の集合の上に定義される
  - 5回のうち表が0回出る
  - 5回のうち表が1回出る
  - 5回のうち表が2回出る
  - 5回のうち表が3回出る
  - 5回のうち表が4回出る
  - 5回のうち表が5回出る

例）5回のうち表が2回出る場合の数

11000

10100

10010

10001

01100

01010

01001

00110

00101

00011

例）5回のコイン投げ (2/3)

- 二項分布は以下の6つの事象の集合の上に定義される
  - 5回のうち表が0回出る・・・ 1通り
  - 5回のうち表が1回出る・・・ 5通り
  - 5回のうち表が2回出る・・・10通り
  - 5回のうち表が3回出る・・・10通り
  - 5回のうち表が4回出る・・・ 5通り
  - 5回のうち表が5回出る・・・ 1通り

例）5回のコイン投げ (3/3)

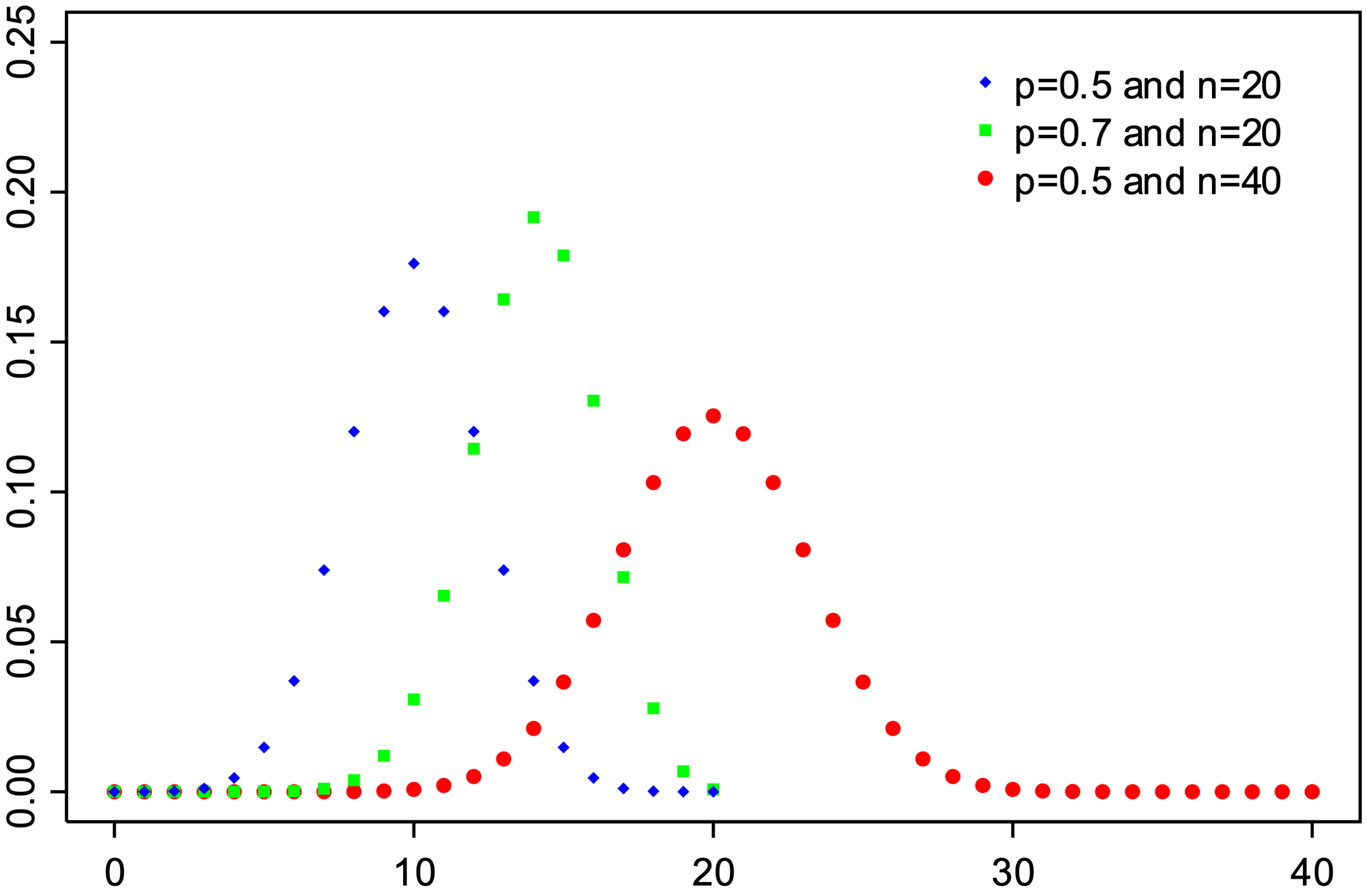
- 二項分布は以下の6つの事象の集合の上に定義される
  - 5回のうち表が0回出る・・・ 1通り　　確率は1/32
  - 5回のうち表が1回出る・・・ 5通り　　確率は5/32
  - 5回のうち表が2回出る・・・10通り　　確率は10/32
  - 5回のうち表が3回出る・・・10通り　　確率は10/32
  - 5回のうち表が4回出る・・・ 5通り　　確率は5/32
  - 5回のうち表が5回出る・・・ 1通り　　確率は1/32

二項分布の確率質量関数

- 確率質量関数 probability mass function; pmf
  - 離散確率変数に、その値をとる確率を対応させる関数

    例：二項分布の場合は、表が出る回数にその確率を対応させる。
- 二項分布の確率質量関数
  - n回のうち表がk回出る確率は以下のように書ける

$$p\left(k;\phi_{\mathrm{head}},n\right)=\frac{n!}{k!\,(n-k)!}{\phi_{\mathrm{head}}}^{k}(1-\phi_{\mathrm{head}})^{n-k}$$



https://en.wikipedia.org/wiki/Binomial_distribution#/media/File:Binomial_distribution_pmf.svg

余談: $\boldsymbol{n} = \mathbf{1}$のときの二項分布の確率質量関数

- 二項分布で$n = 1$とする（kは0か1のどちらかになる）

$$\phi_{\mathrm{head}}{}^{k}(1 - \phi_{\mathrm{head}})^{1-k}$$

- こういうの、どっかで見た！
  - 対数をとってマイナスをつければ、ロジスティック回帰で使っていたクロスエントロピーになる。

$$-k \log \phi_{\mathrm{head}} - (1 - k) \log(1 - \phi_{\mathrm{head}})$$

独立同分布の仮定
(independently and identically distributed; i.i.d.)

$$\frac{n!}{k!\,(n-k)!}{\phi_{\mathrm{head}}}^{k}(1-\phi_{\mathrm{head}})^{n-k}$$

- n回のうち表がk回出る確率を上式で表すとき、以下のことを仮定している
  - 表が出る確率は履歴に左右されない
    - 試行の独立性の仮定
  - 表が出る確率は同じままである
    - 確率分布の同一性の仮定

問題2-1

- $\phi_{\mathrm{head}} = 0.6$であるコインを3回投げたとき、
  - 表が1回も出ない確率
  - 表が1回だけ出る確率
  - 表がちょうど2回出る確率
  - 3回とも表が出る確率

を、それぞれ求めよ。ただしi.i.d.を仮定する。

最尤推定

統計モデリングの問い

「どのモデルが、観測されたデータを、
一番うまく説明してくれるか？」

- 「どのモデル」：どの範囲で選ぶか？
  例：いろいろありうる二項分布のなかから選ぶ
- 「一番うまく説明してくれる」：どういう意味？

二項分布の場合の統計モデリングの問い

「ϕ_{head}がとりうる0~1の値のうちどの値が、

観測データを一番うまく説明してくれるか？」

- ϕ_{head}を決めれば、無数にありうる二項分布のうちのひとつを選んだことになる
  - 裏が出る確率はϕ_{head}から自動的に決まる。
  - nの値は観測データから決まる。
- どうやってϕ_{head}の値を決める？

パラメータを使ってデータの確率を書く

- 「コインをn回投げて、表がk回出る確率はいくら？」
  - この問いにϕ_{head}を使って答えると…

$$\frac{n!}{k!\,(n-k)!}{\phi_{\mathrm{head}}}^{k}(1-\phi_{\mathrm{head}})^{n-k}$$

  - パラメータを使えば観測データが生成される確率を表せる
  - あとはϕ_{head}のところに値をあてはめればよいが・・・

しかしパラメータの値は分かっていない

- もしコインについてϕ_{head}が分かっていれば…
  - 前のスライドの式より、コインをn回投げたとき表がk回出る確率を計算できる
- しかし現実には…
  - コインを実際にn回投げて表が何回出るかを調べることはできるものの、ϕ_{head}を直接知る方法は無い

データからパラメータを推定する

- 目の前にあるこのコインを投げて表が出る確率がいくらかなんて誰も知らない（だから統計的にモデリングしている）
- しかし、実際にコインを投げて表が出た回数から、逆にその確率を推定することはできる
  - これが統計モデリングにおける二項分布の使い方
  - 観測データからパラメータを推定estimateする

問題2-2

- コインを100回投げたら、表が52回出たとする
- このコインの表が出る確率ϕ_{head}はいくらか、推定してみよう

問題2-2の解答例

- 100回のうち表が52回出たのだから・・・

$$\phi_{\mathrm{head}} = \frac{52}{100} = 0.52$$

- なぜこの割り算でいいの？
  - 普通こう計算すると思うが、なぜこう計算するのでいいの？

解答例への理屈づけ

$$\frac{100!}{52!\,(100-52)!}{\phi_{\mathrm{head}}}^{52}(1-\phi_{\mathrm{head}})^{100-52}$$

- この式は、表が出る確率がϕ_{head}であるコインを100回投げたときに表が52回出る確率を表している
- この式を ϕ_{head}の関数とみなす
- そして、この関数の値を最大にするϕ_{head}はいくらかを求める
  - 次スライドへ

$$f(x) = \frac{100!}{52!\,(100-52)!} x^{52}(1-x)^{100-52}$$

$$\frac{df(x)}{dx} = \frac{100!}{52!\,(100-52)!} \{52x^{51}(1-x)^{48} - 48x^{52}(1-x)^{47}\}$$

$$= \frac{100!}{52!\,(100-52)!} x^{51}(1-x)^{47}\{52(1-x) - 48x\}$$

$\frac{df(x)}{dx} = 0$とおくと

$$52(1-x) - 48x = 0$$

$$\therefore x = \frac{52}{100}$$

尤度 likelihood

$$\frac{100!}{52!\,(100-52)!}{\phi_{\mathrm{head}}}^{52}(1-\phi_{\mathrm{head}})^{100-52}$$

- 上の例のように、確率分布のパラメータの関数として表された観測データの生成確率のことを、尤度と呼ぶ
- 尤度はパラメータの関数である
  - 観測データから分かる値（上の例では100や52のこと）は、定数。
  - 尤度は観測データの関数ではなく、パラメータの関数。

最尤推定 maximum likelihood estimation

- 観測データの尤度を最大化することによってモデルのパラメータを推定することを最尤推定と呼ぶ
- 尤度はモデルパラメータの関数
  - パラメータの値を変えると尤度が変わる、ということ
  - 観測データは既知なので定数、つまり固定された値

ベイズ的な考え方

問題2-2の解答例（再び）

- 100回のうち表が52回出たのだから・・・

$$\phi_{\mathrm{head}} = \frac{52}{100} = 0.52$$

- しかし…この求め方で本当にいいのか？
  - ダメだとすればその理由は？

この解答例の問題点

- 表が出る確率ϕ_{head}を、その100回のコイン投げのデータだけで決めてしまっていいのか？
  - たまたまその100回では表が52回出ただけではないのか？

ベイズ的な考え方

- 表が出る確率ϕ_{head}を決めること自体に不確かさがある、と考える！
- つまり、ϕ_{head}が0~1の範囲のいくらの値をとるかも、確率的に決まる、と考える
  - パラメータの値自体が確率的に決まると考える
  - これがベイズ統計の考え方

ベイズは（最低）２段構え

- コイン投げで表が何回出るかは確率的に決まる
- それだけでなく、表が出る確率も確率的に決まる
- こう考えるのが、ベイズ的な考え方

パラメータの確率分布って？

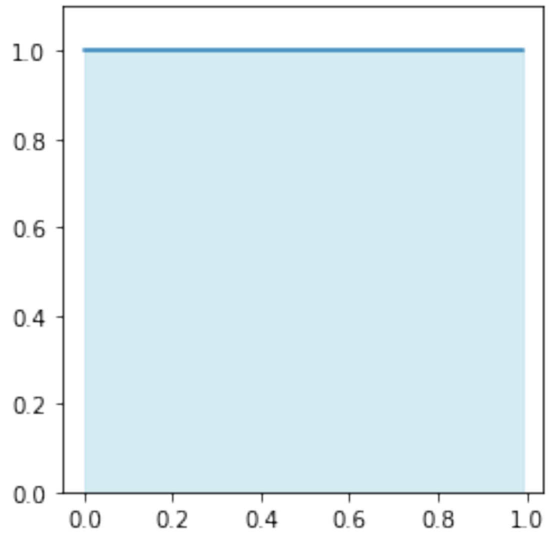
- 表が出る確率ϕ_{head}は様々でありうる
  - 0.01, 0.1, 0.3, 0.6, 0.98等々・・・0以上1以下なら何でもOK
- とびとびの値ではなく連続的な値
  - 表が出る確率の値が何通りあるかなんて言えない
    - 強いて言えば、無限通りある。
- つまり0~1の連続的な値の確率を考える必要がある

連続確率分布

- 連続的な値の確率を考えるときは連続確率分布を使う
  - 連続分布は確率密度関数probability density function; pdfで表すことが多い
  - 確率密度関数は定義域全域で積分すると1になる
  - 累積密度関数cumulative distribution function;cdfも使う
    - ざっくり言えば、cdfを微分するとpdfになる
      - 正確には、まずcdfを定義し、次にpdfを導けるときは導く、という順

一様分布

- 連続確率分布のうち最も単純な分布と言っていいかも
  - 定義域はいろいろ。
    - 下図は[0,1]が定義域の場合の一様分布の確率密度関数。
    - 積分すると（＝グラフの下側の面積が）1になる。



復習：ベイズ則

$$p(z|x) \propto p(x|z)p(z)$$

- データを観測すると仮説の確率が変わる
  - ある仮説 z が成り立つ確率は・・・
  - その仮説が成り立つとき尤もらしくなるデータ x が観測されると・・・
  - 高くなる

ベイズ則をベイズ的に使う

- 表が出る確率がいくらかということ自体の<u>不確かさ</u>
  - これを考えるのがベイズ的な考え方
- ベイズ的な考え方をベイズ則できちんと表す
  - コインの表が出たり裏が出ることを観測することによって、この<u>不確かさuncertainty</u>がどのように変化するか。
  - ベイズ則を使うと、これを数学的にきちんと記述できる。

コインの投げのベイズ的モデリング

$$p(\phi_{\mathrm{head}}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\phi_{\mathrm{head}})p(\phi_{\mathrm{head}})$$

- コイン投げの結果$\boldsymbol{x}$を観測することで、コインの表が出る確率ϕ_{head}としてどういう値が<u>よりありえそうか</u>の、分布が変わりますよ、という式
  - $\boldsymbol{x}$はたくさんのコイン投げの結果をまとめて表す記号
    - $\boldsymbol{x} = (x_1 = 1, x_2 = 0, x_3 = 1, \dots)$: 表、裏、表、・・・
- ϕ_{head}が従う分布についてどういう分布からスタートする？

一様分布からスタートするとする

- コインの表が出る確率がいくらかなんて全く分からないという状態からスタートすることにする
  - そこで、 ϕ_{head}が[0,1]の区間のどの値をとるかについては、一様分布を仮定する
  - つまり$p(\phi_{\mathrm{head}}) = 1$
    - どの値も同じようにありえそう、という意味

1回コインを投げたら表だったとする

$$p(\phi_{\mathrm{head}}|x_1 = 1) \propto p(x_1 = 1|\phi_{\mathrm{head}})p(\phi_{\mathrm{head}})$$

右辺は

$$p(x_1 = 1|\phi_{\mathrm{head}}) = \phi_{\mathrm{head}}$$
$$p(\phi_{\mathrm{head}}) = 1$$

よって

$$p(\phi_{\mathrm{head}}|x_1 = 1) \propto \phi_{\mathrm{head}}$$

確率密度関数の正規化

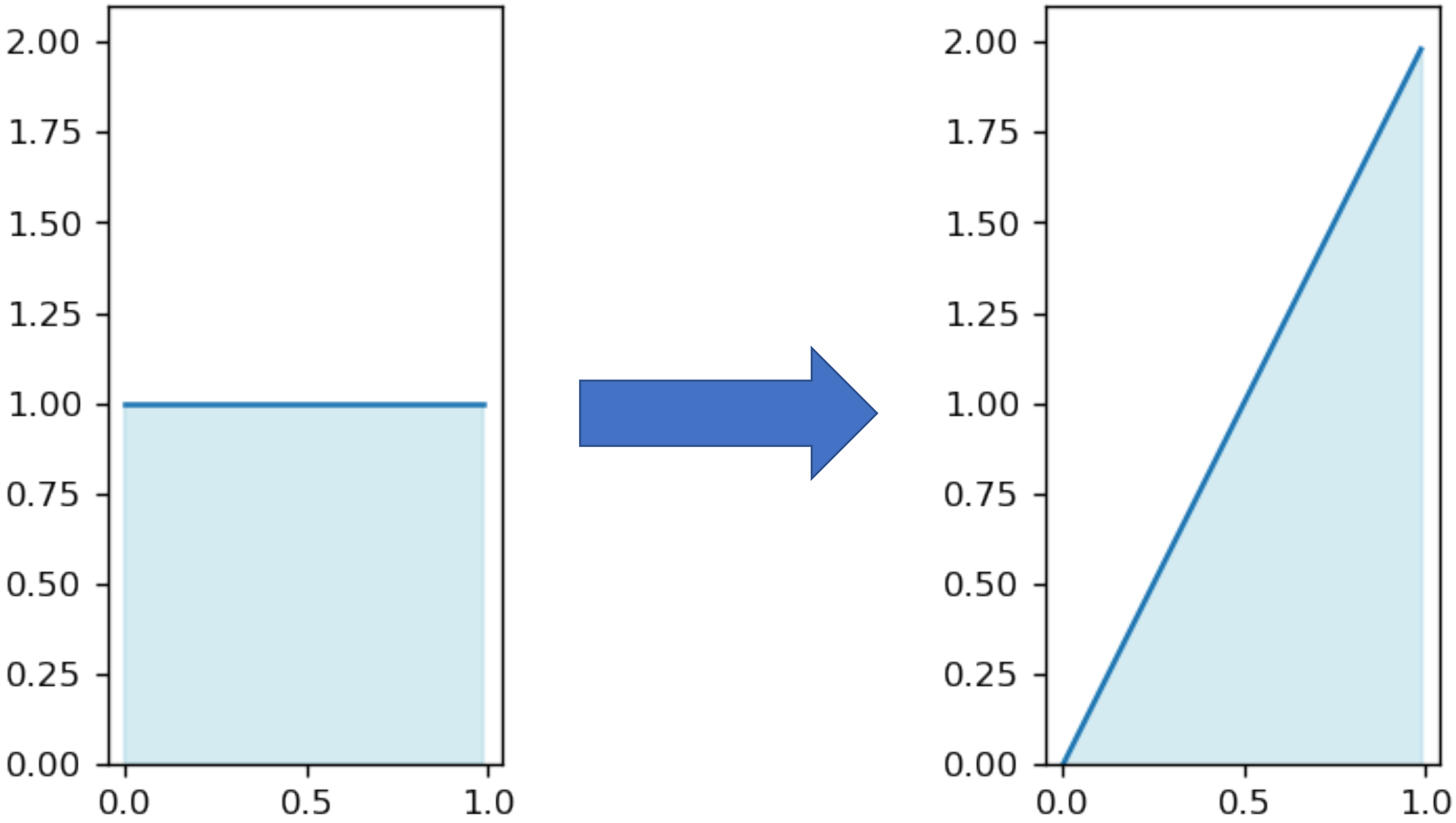
確率密度関数は定義域全域で積分すると1

$p(\phi_{\text{head}}|x_1 = 1) \propto \phi_{\text{head}}$を満たすから

$$\int_0^1 \phi_{\text{head}} d\phi_{\text{head}} = \left[\frac{1}{2}\phi_{\text{head}}{}^2\right]_0^1 = \frac{1}{2}$$

よって、積分して1になるためには

$$p(\phi_{\text{head}}|x_1 = 1) = 2\phi_{\text{head}}$$



2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0

もう1回コインを投げたら裏だったとする

$$p(\phi_{\text{head}} | x_1 = 1, x_2 = 0) \propto p(x_1 = 1, x_2 = 0 | \phi_{\text{head}}) p(\phi_{\text{head}})$$

右辺は

$$p(x_1 = 1, x_2 = 0 | \phi_{\text{head}}) = \phi_{\text{head}}(1 - \phi_{\text{head}})$$
$$p(\phi_{\text{head}}) = 1$$

よって、積分して1になるためには

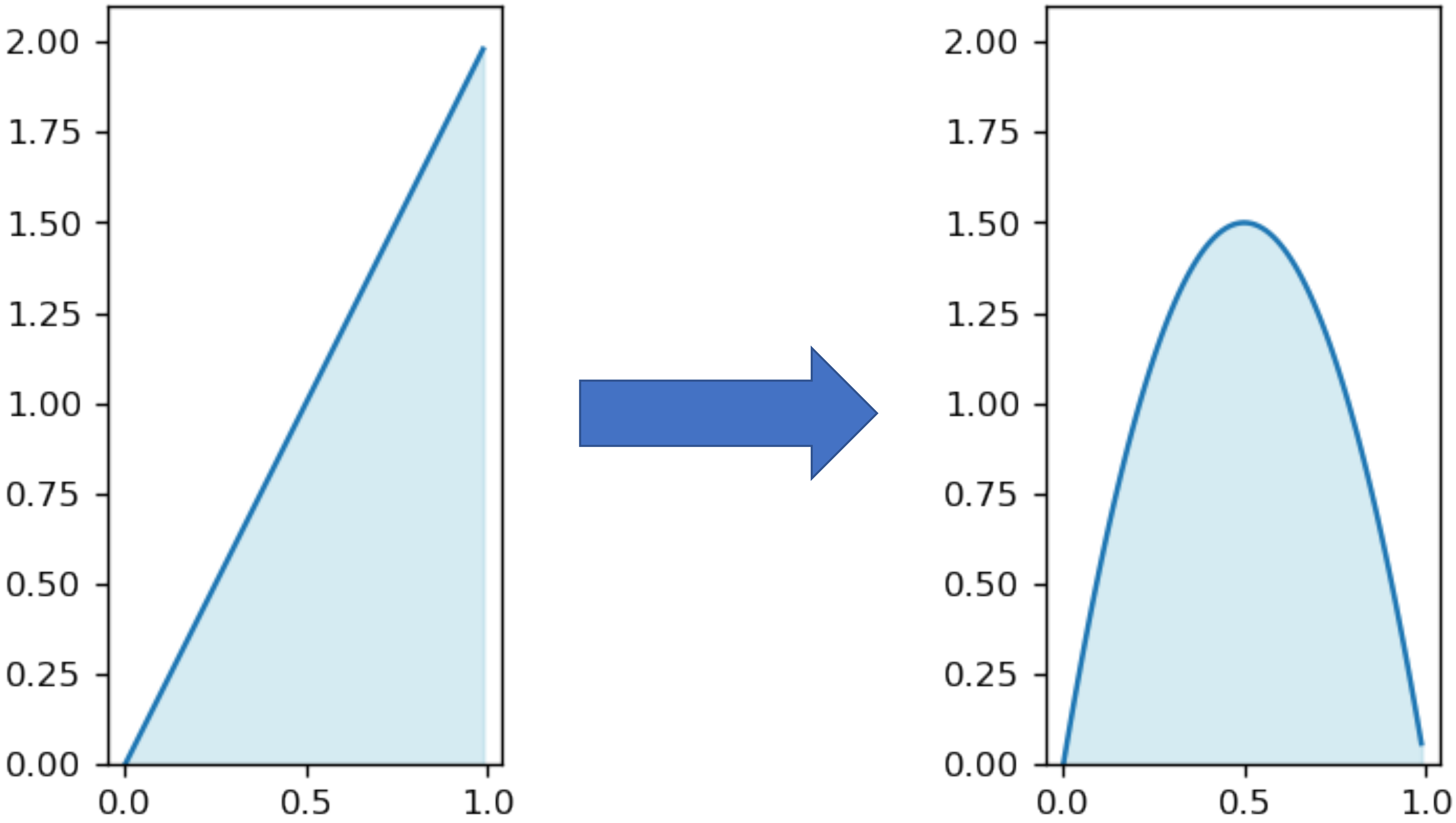
$$p(\phi_{\text{head}} | x_1 = 1, x_2 = 0) \propto \phi_{\text{head}}(1 - \phi_{\text{head}})$$

確率密度関数の正規化

$$\int_0^1 \phi_{\text{head}}(1-\phi_{\text{head}})d\phi_{\text{head}} = \left[\frac{1}{2}{\phi_{\text{head}}}^2 - \frac{1}{3}{\phi_{\text{head}}}^3\right]_0^1 = \frac{1}{6}$$

よって

$$p(\phi_{\text{head}} | x_1 = 1, x_2 = 0) = 6\phi_{\text{head}}(1-\phi_{\text{head}})$$


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0

もう1回コインを投げたら表だったとする

$$p(\phi_{\mathrm{head}} | x_1 = 1, x_2 = 0, x_3 = 1) \propto p(x_1 = 1, x_2 = 0, x_3 = 1 | \phi_{\mathrm{head}}) p(\phi_{\mathrm{head}})$$

右辺は

$$p(x_1 = 1, x_2 = 0, x_3 = 1 | \phi_{\mathrm{head}}) = {\phi_{\mathrm{head}}}^2 (1 - \phi_{\mathrm{head}})$$

$$p(\phi_{\mathrm{head}}) = 1$$

よって

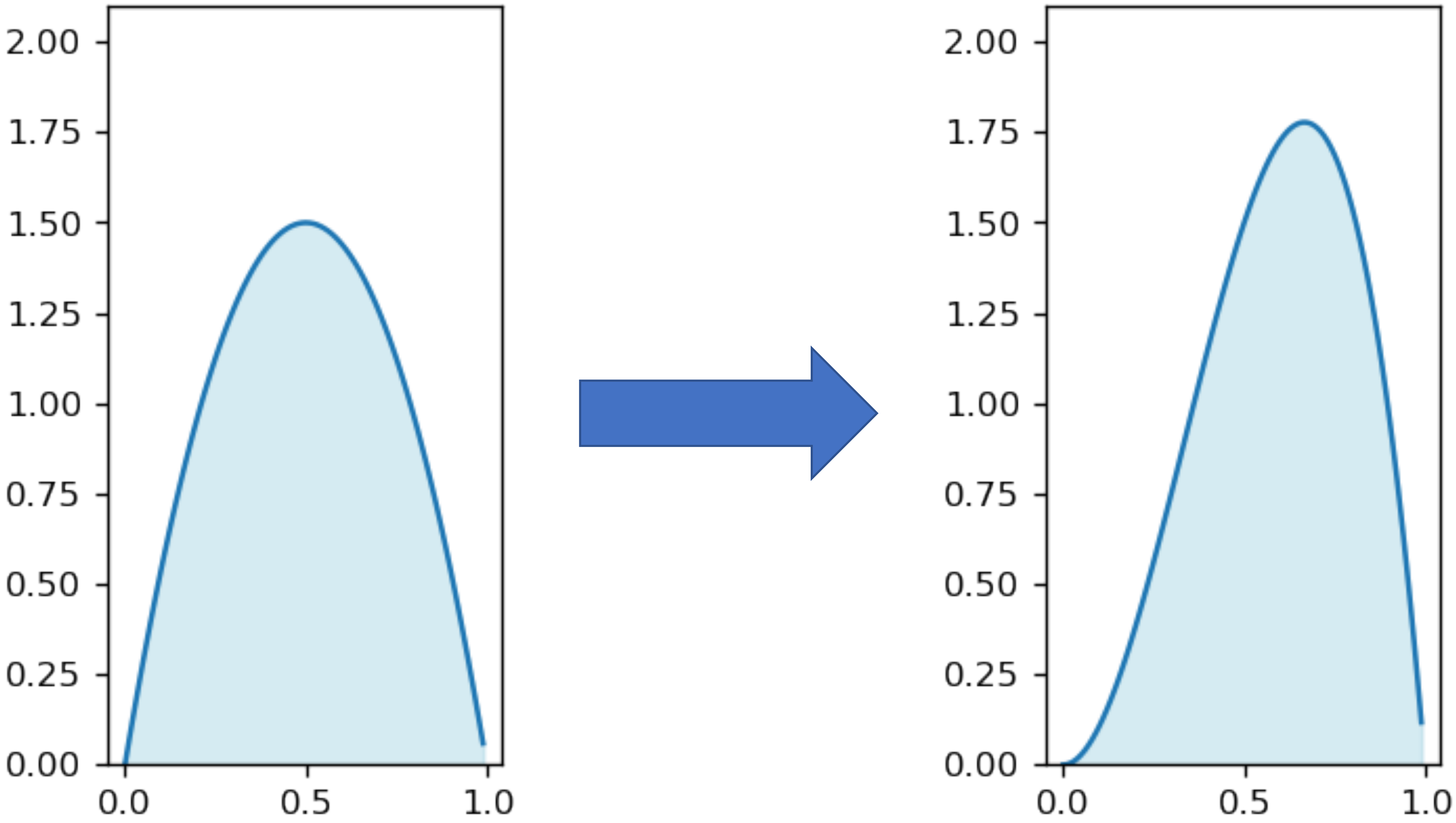
$$p(\phi_{\mathrm{head}} | x_1 = 1, x_2 = 0, x_3 = 1) \propto {\phi_{\mathrm{head}}}^2 (1 - \phi_{\mathrm{head}})$$

確率密度関数の正規化

$$\int_0^1 {\phi_{\text{head}}}^2(1-\phi_{\text{head}})d\phi_{\text{head}} = \left[\frac{1}{3}{\phi_{\text{head}}}^3 - \frac{1}{4}{\phi_{\text{head}}}^4\right]_0^1 = \frac{1}{12}$$

よって、積分して1になるためには

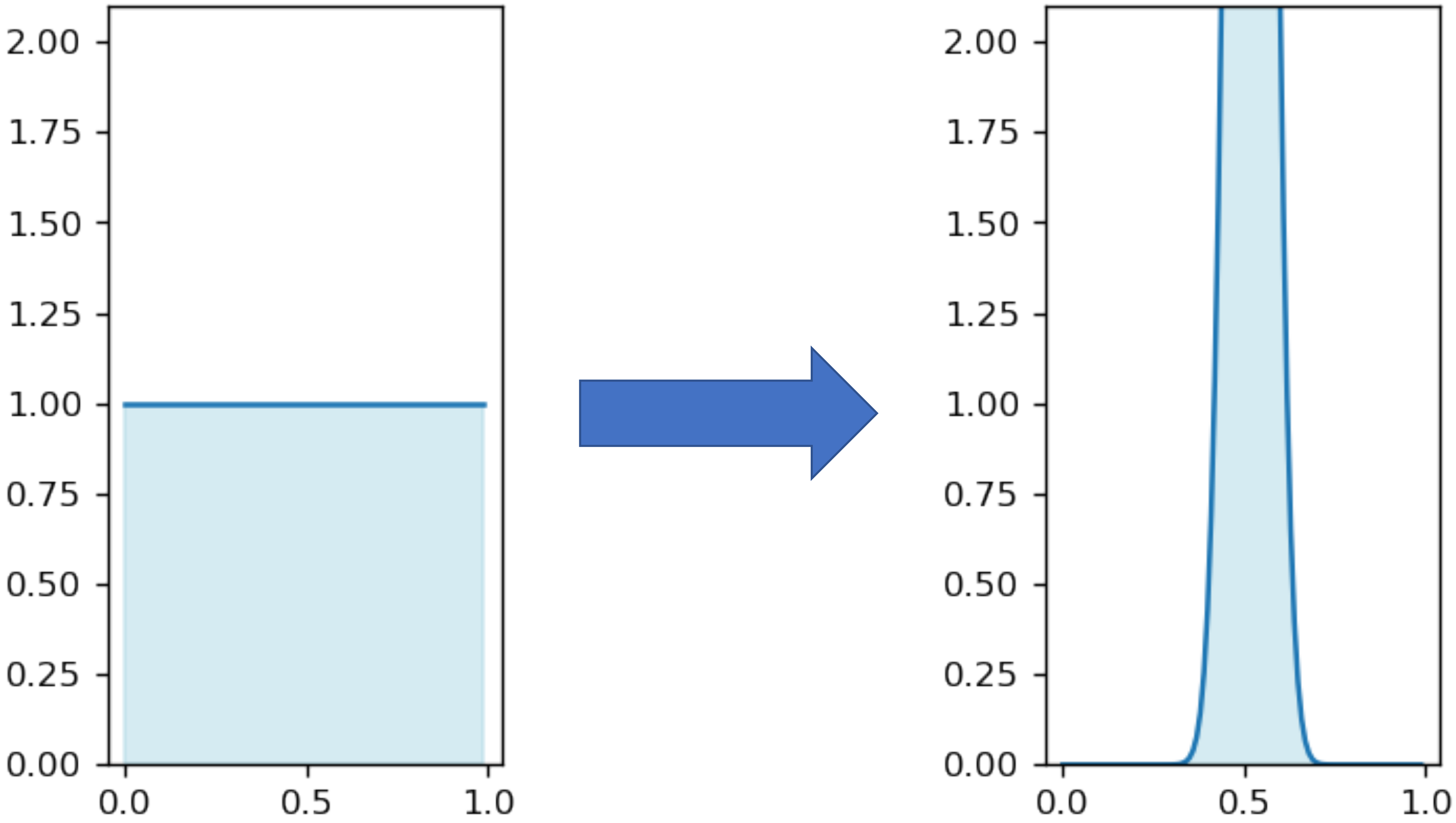
$$p\left(\phi_{\text{head}} \middle| x_1 = 1, x_2 = 0, x_3 = 1\right) = 12{\phi_{\text{head}}}^2(1-\phi_{\text{head}})$$


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0

合計$\boldsymbol{n}$回投げて表が$\boldsymbol{k}$回出た

$$p(\phi_{\mathrm{head}}|\boldsymbol{x}) = \frac{n!}{k!\,(n-k)!}{\phi_{\mathrm{head}}}^{k}(1-\phi_{\mathrm{head}})^{n-k}$$

例えば、合計100回投げて表が52回出たとすると・・・


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0

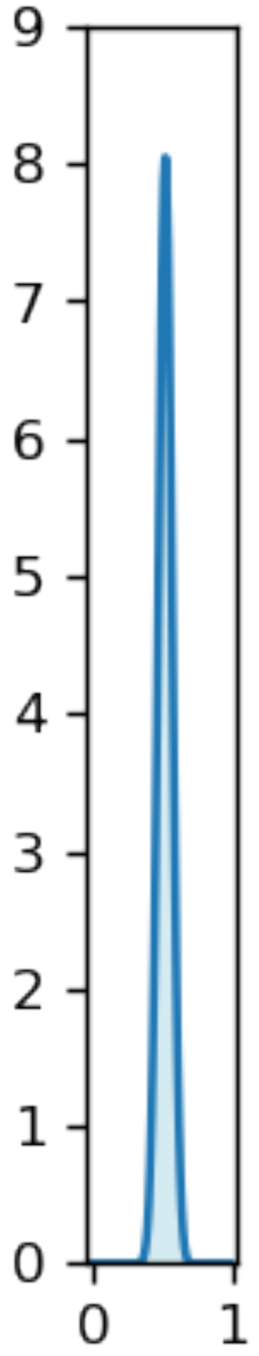
合計$\boldsymbol{n}$回投げて表が$\boldsymbol{k}$回出た（再）

$$p\left(\phi_{\mathrm{head}} \middle| \boldsymbol{x}\right) = \frac{n!}{k!\,(n-k)!}\,{\phi_{\mathrm{head}}}^{k}(1-\phi_{\mathrm{head}})^{n-k}$$

- 「これって、二項分布のpdfと同じでは？」
- 確かに式の形は同じ。ベイズ則$p\left(\phi_{\mathrm{head}} \middle| \boldsymbol{x}\right) \propto p\left(\boldsymbol{x} \middle| \phi_{\mathrm{head}}\right) p\left(\phi_{\mathrm{head}}\right)$においてて、今回は$\phi_{\mathrm{head}}$がいくらであっても$p\left(\phi_{\mathrm{head}}\right) = 1$と仮定している（一様分布を仮定している）から、二項分布のpdfと同じ式になる。
- だが、<u>意味は全く違う</u>。これはϕ_{head}の関数。kの関数ではない。

最初の解答例との違い

- 「合計100回投げて表が52回出た」という出題に、最初の解答例では52/100と答えを求めていた
- 今回は、連続的な関数（右図）で答えている
  - 表が出る確率として考えられる0から1の値について、それぞれがどのくらいありえそうかで、答えている。
    - 答えをひとつの値に決めていない、とも言える。
  - これがベイズ推定。



ベイズ推定 Bayesian inference

- モデルのパラメータの値を推定するという問題に
- 最尤推定のようにひとつの値によってではなく
- 様々な値がそれぞれどのくらいありえそうかで答える
- つまり、モデルパラメータのとりうる値の集合の上に定義された確率分布によって答える
  - これがベイズ推定。

ベイズ推論 Bayesian inference

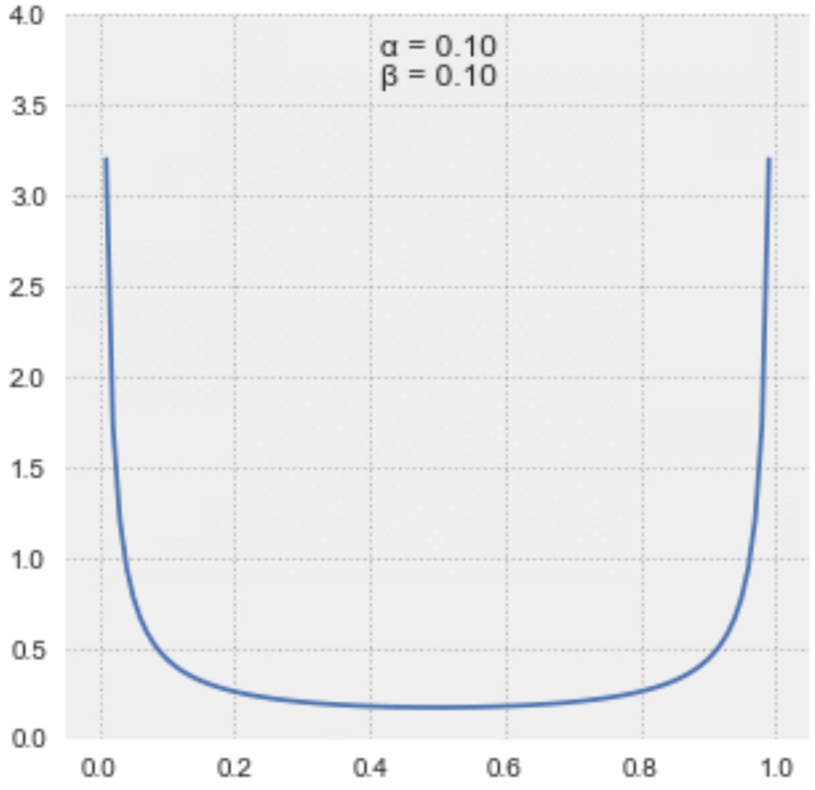
- ベイズ的な考え方では、パラメータの値を推定するという問題に、ひとつの値で答えず、分布で答える。
- そのため、日本語で「ベイズ推定」という言い方をせずに、「ベイズ推論」という言い方をすることで、この答え方の違いを、字面の違いとして表すことがある。
  - 深層学習の世界では、学習済みモデルを使った前向き計算のことを、推論 inferenceと呼んでいたりするので、混乱しないようにしましょう。

ベータ分布 beta distribution

- 上の説明に出てきた連続分布をベータ分布と呼ぶ
- ベータ分布は2つのパラメータa, bを持つ
- 確率密度関数は以下の式で与えられる

$$p(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$$

ベータ分布の密度関数



https://en.wikipedia.org/wiki/File:PDF_of_the_Beta_distribution.gif

コイン投げにおけるベイズ推定

- 表が出る確率の値は無数にありうる
- その無数にありうる値それぞれがどのくらいありえそうかを、ベータ分布で表わすことができる
- 自分の持っているコインについて
  - 例えば一様分布からスタートし・・・
  - コインを投げながらベータ分布をアップデートしていく

事前分布と事後分布

- 事前分布 prior distribution
- 事後分布 posterior distribution
    - どちらもすでに登場しています
    - 今日の話では、一様分布が事前分布で、ベイズ則を適用して得たベータ分布が事後分布。

$$p(\phi_{\text{head}}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\phi_{\text{head}})p(\phi_{\text{head}})$$

ベイズ推論におけるベイズ則

- ベイズ推論で使うベイズ則は、単に条件付き確率どうしの関係を表すだけの式ではない

（事後分布）∝（尤度）✖（事前分布）

- 上の式が、ベイズ推論におけるベイズ則の意味

課題2

1. ϕの関数$f(\phi) = \phi^3(1-\phi)^2$を$0 \leq \phi \leq 1$の範囲で積分せよ。
2. 積分した結果が1になるようにするためには、$f(\phi)$にいくらを掛けておけばよかったか？