OLIS-早稲田大学保険フォーラム
アクチュアリーとデータサイエンス

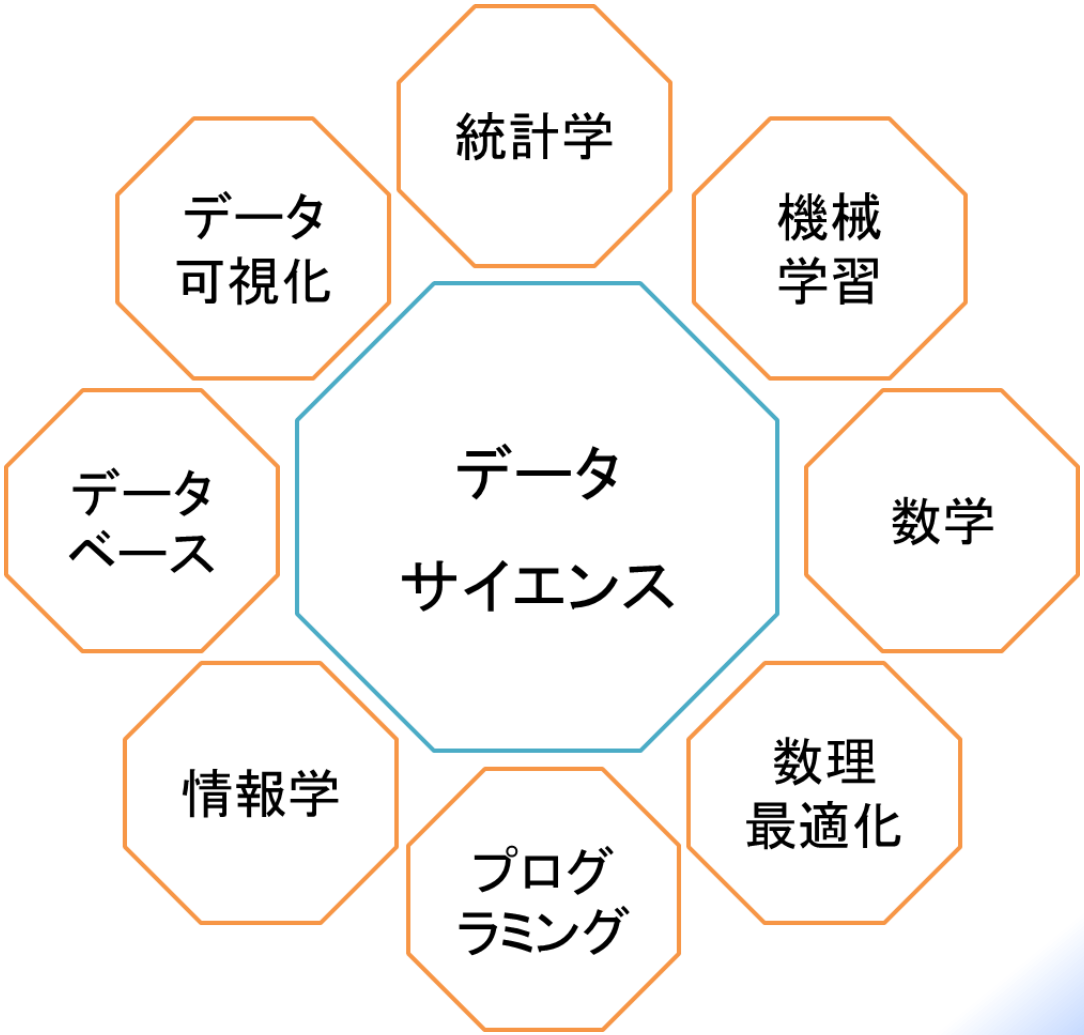
データサイエンスを学ぶ

2021年11月13日

早稲田大学 大学院会計研究科　野村俊一

データサイエンスは幅広い

- データサイエンスは
  データを扱う分野の
  複合で構成されている
- 本講演では機械学習
  と統計学の考え方の
  違い（個人的見解）を
  一例で紹介した後、
  機械学習の基本である
  ベイズモデルを簡単に
  解説する



スポーツ選手の誕生月

下表の各スポーツの誕生月別選手数のデータについて、
『選手の誕生月が概ね均等に分布している』と思われるか
どうかをスポーツ種目ごとに自分の主観で答えてください

誕生月	4月	5月	6月	7月	8月	9月	10月	11月	12月	1月	2月	3月
野球	90	99	95	93	87	66	58	47	60	41	37	40
柔道	23	35	28	31	29	17	25	17	16	17	11	17
ボクシング	21	17	15	25	29	25	24	22	20	14	23	22
ジョッキー	7	3	9	12	12	13	15	21	11	11	23	28

統計学　「表のデータが『均等な分布から得られた』という仮説が確率的に妥当かどうか判定したい（仮説検定）」

機械学習「分布が均等かどうかは考えずに、新たな選手の誕生月を予測したい」（統計学以上に予測を重視する）

一様性の検定

データが各区分に均等に分布しているかどうかを、次の一様性の検定（カイ二乗適合度検定）により判定する

一様性の検定

k 個の区分の度数がそれぞれ $n_1, n_2, \ldots, n_k$（合計 n）のとき、帰無仮説：各区分の所属確率は均等、対立仮説：左記以外の有意水準5%での仮説検定は、次のカイ二乗検定統計量

$$X = \sum_{i=1}^{k} \frac{(n_i - n/k)^2}{n/k}$$ （n/k は各区分の度数の期待値）

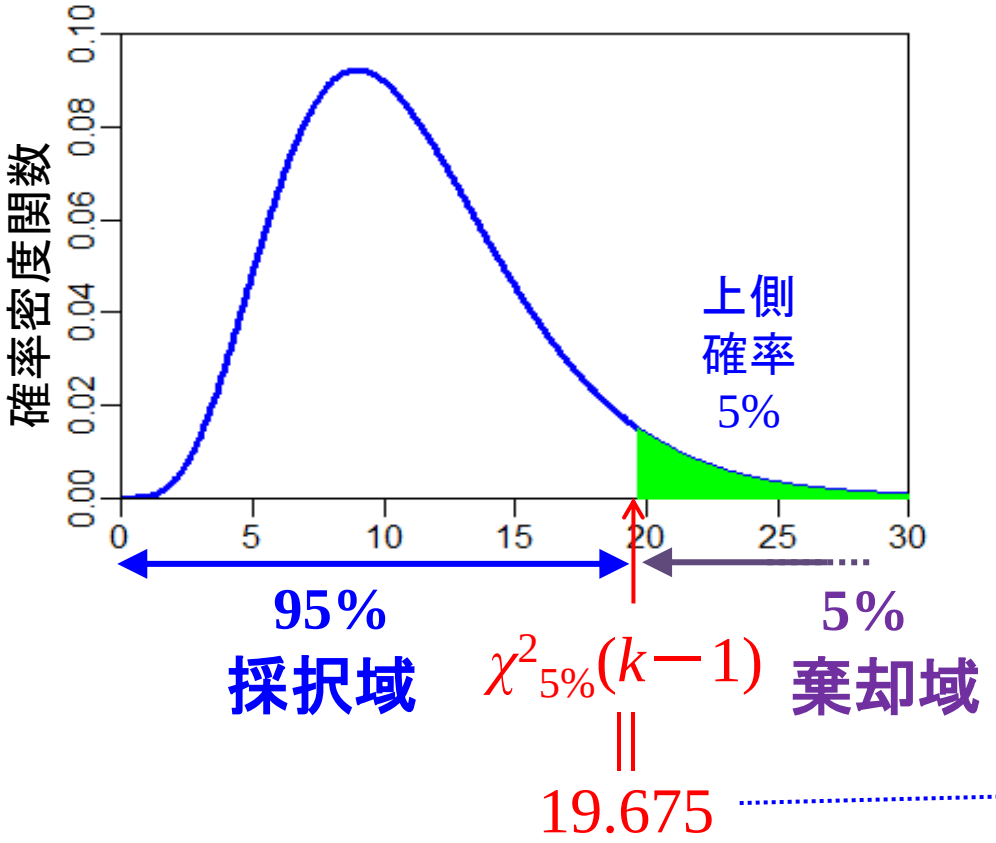
が棄却域 $X > \chi^2_{5\%}(k-1)$ に入れば棄却（均等でないと言える）、採択域に入れば採択（均等でないとは言えない）と結論する

※ $\chi^2_{5\%}(k-1)$ は自由度 $k-1$ の χ^2 分布の上側5%点

χ^2（カイ二乗）分布の上側5%点

前頁のカイ二乗検定統計量 X は、近似的に自由度 $k-1$ の χ^2 分布に従い、各区分の度数が不均等なほど大きくなる

$k=12$ のとき X が従う自由度 $k-1=11$ の χ^2 分布の上側5%点 $\chi^2_{5\%}(k-1)$



$k-1$	$\chi^2_{5\%}(k-1)$
1	3.841
2	5.991
3	7.815
4	9.488
5	11.070
6	12.592
7	14.067
8	15.507
9	16.919
10	18.307
11	19.675

スポーツ選手の誕生月（一様性の検定）

種目ごとに一様性の検定を行った結果は下表の通り

（棄却域：$X > \chi^2_{5\%}(12-1) = 19.675$ に入れば棄却（均等でない）、
採択域：$X \leq 19.675$ に入れば採択（均等でないと言えない）という結論となる）

誕生月	4月	5月	6月	7月	8月	9月	10月	11月	12月	1月	2月	3月	計	カイ二乗検定統計量	結論
野球	90	99	95	93	87	66	58	47	60	41	37	40	813	92.13653137	棄却
柔道	23	35	28	31	29	17	25	17	16	17	11	17	266	27.14285714	棄却
ボクシング	21	17	15	25	29	25	24	22	20	14	23	22	257	9.848249027	採択
ジョッキー	7	3	9	12	12	13	15	21	11	11	23	28	165	39.87272727	棄却

スポーツ選手の誕生月（予測）

- 先ほどのデータから、各種目の新たな選手の誕生月を予測（各月の確率を相対度数などで評価）することができる
- データが少ない種目の場合、相対度数をそのまま確率にして予測すると極端な予測になりやすい

⇒ 極端な予測を回避する1つの方策として、
ベイズモデルを用いた予測方法を紹介する

誕生月データ例

誕生月	4月	5月	6月	7月	8月	9月	10月	11月	12月	1月	2月	3月	計
度数	8	6	7	6	4	4	5	4	2	0	3	1	50
相対度数	0.16	0.12	0.14	0.12	0.08	0.08	0.10	0.08	0.04	0.00	0.06	0.02	1.00

度数0の月が誕生月となる確率を0と予測して良いのか↑

ベイズモデル入門:コイン投げシミュレーションゲーム

- 統計ソフトウェアRを用いて次のゲームを行います

Step. 1 統計ソフトを用いて 0 以上 1 以下の一様乱数を生成して p とおく(皆さんには知らされません)

Step. 2 表が出る確率 p のコインを 2 回投げるシミュレーションを乱数で行い、表が出た回数 X を画面に表示する

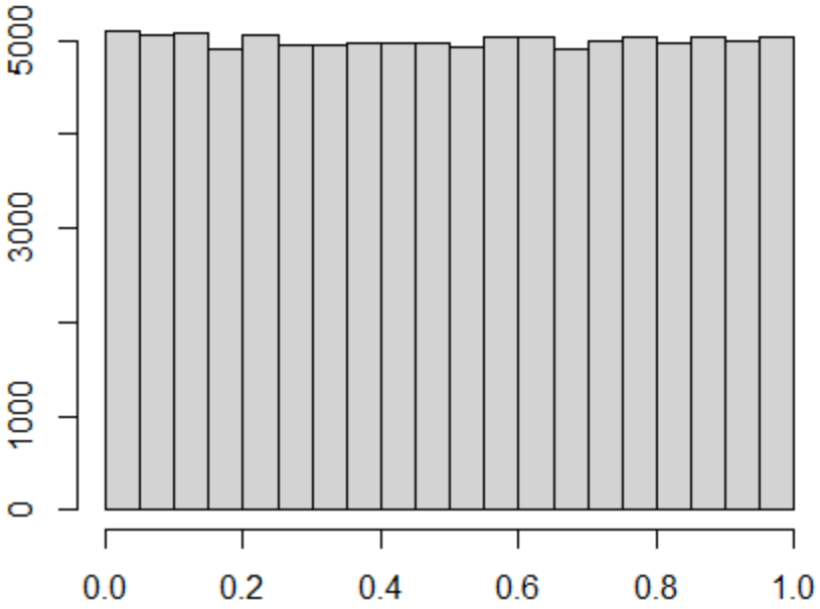
Step. 3 表が出る確率 p の値を皆さんで予想してください

```
> # Step.1 一様乱数pを生成
> p = runif(1)
>
> # Step.2 表が出る確率pの2回のコイン投げで表が出た回数Xを表示
> (X = rbinom(1,2,p))
[1] 1
>
> # Step.3 pの値を予想してください
> p
```

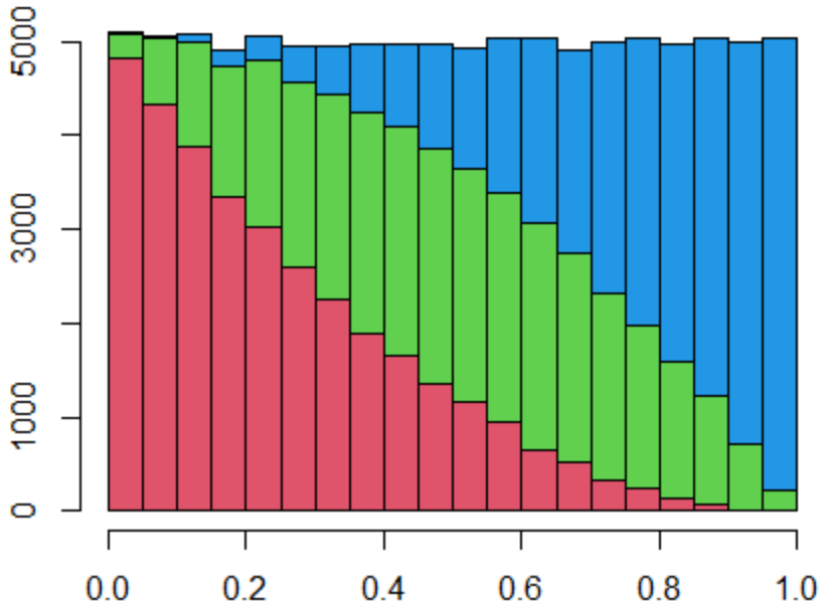
シミュレーションによる考察

- ゲームを10万回分シミュレートして、p と X の実現値の分布を確認する

Step. 1で生成された p の値のヒストグラム



Step. 2で生成された X の値で色分け



- $X = 2$ となった p（**33,241個、平均0.75**）
- $X = 1$ となった p（**33,478個、平均0.50**）
- $X = 0$ となった p（**33,281個、平均0.25**）

最良の推定値

定理

$X = x$ となった p のシミュレーション値 $p_1, \ldots, p_N$ に対し

平均二乗誤差 $\frac{1}{N}\sum_{i=1}^{N}(\hat{p} - p_i)^2$ を最小にする $\hat{p}$ は

$\hat{p} = \bar{p} = \frac{1}{N}\sum_{i=1}^{n} p_i$（シミュレーション値の平均値）である

証明：平均二乗誤差 $\frac{1}{N}\sum_{i=1}^{N}(\hat{p} - p_i)^2$ を変形すると

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{p} - p_i)^2 = \frac{1}{N}\sum_{i=1}^{N}\{(\hat{p} - \bar{p}) - (p_i - \bar{p})\}^2$$

$$= (\hat{p} - \bar{p})^2 - \frac{2}{N}(\hat{p} - \bar{p})\sum_{i=1}^{N}(p_i - \bar{p}) + \frac{1}{N}\sum_{i=1}^{N}(p_i - \bar{p})^2$$

$$= (\hat{p} - \bar{p})^2 + \frac{1}{N}\sum_{i=1}^{N}(p_i - \bar{p})^2 \geq \frac{1}{N}\sum_{i=1}^{N}(p_i - \bar{p})^2$$

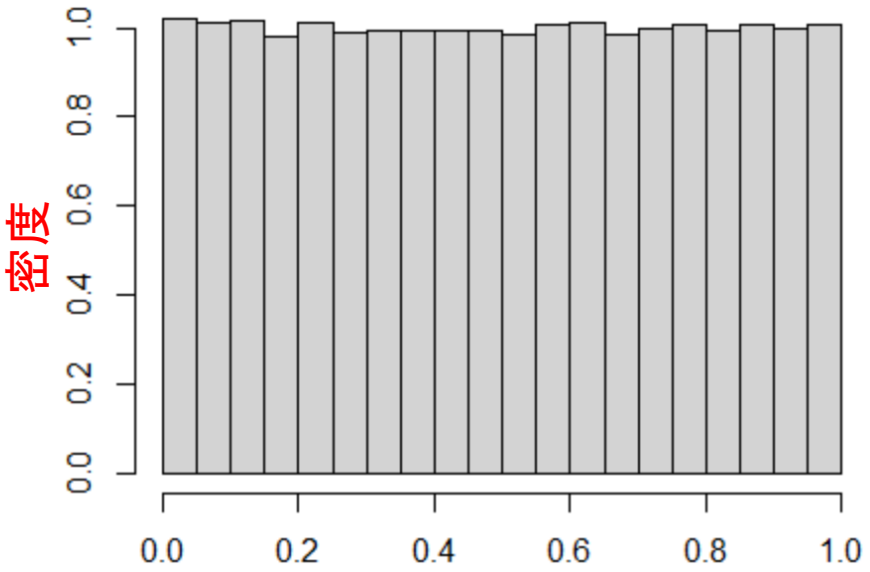
となるので、$\hat{p} = \bar{p}$ のとき平均二乗誤差が最小となる

シミュレート数を無限に増やすと・・・

- 生成数を増やし階級幅を狭くすると、ヒストグラム※は乱数を生成する分布の**確率密度関数**へと近づいていく

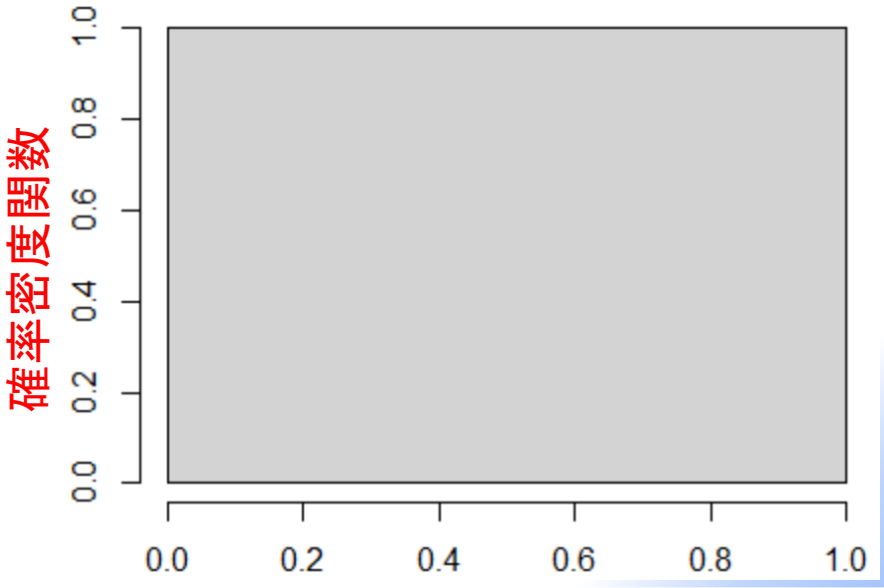
※ 柱の高さを **密度 = 相対度数 ÷ 柱の幅** とすることで、柱の総面積（柱の面積 = 柱の幅 × 高さ = 相対度数 の総和）が 1（全確率）に固定される

Step. 1で生成された p の値のヒストグラム



p の生成数を無限にした極限

p の確率密度関数 $f(p) = 1\ (0 \leq p \leq 1)$



シミュレート数を無限に増やすと・・・

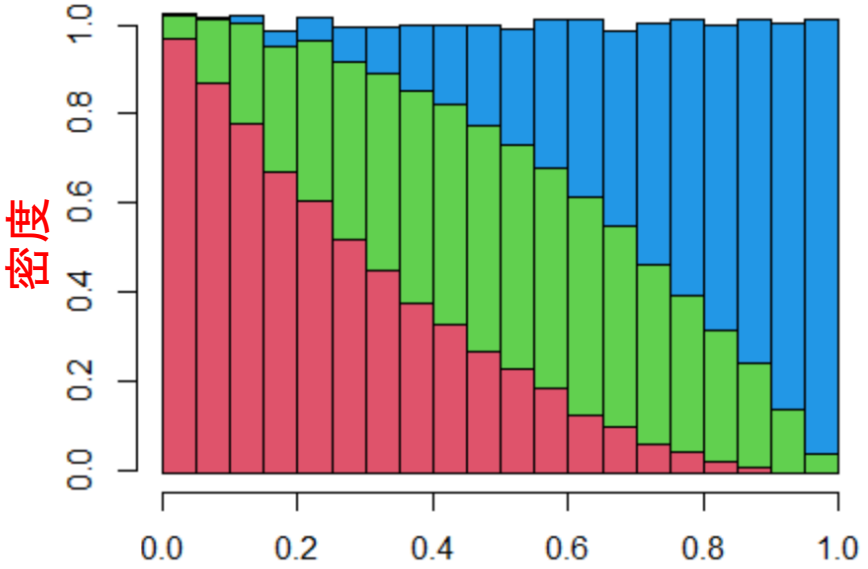
- p の値に対する X の生成確率（条件付き確率）:

$$\mathrm{P}(X=0\,|\,p)=(1-p)^2,\ \mathrm{P}(X=1\,|\,p)=2p(1-p),\ \mathrm{P}(X=2\,|\,p)=p^2$$

- 上式の p に関する期待値で $\mathrm{P}(X=x)$（周辺確率）を得る:

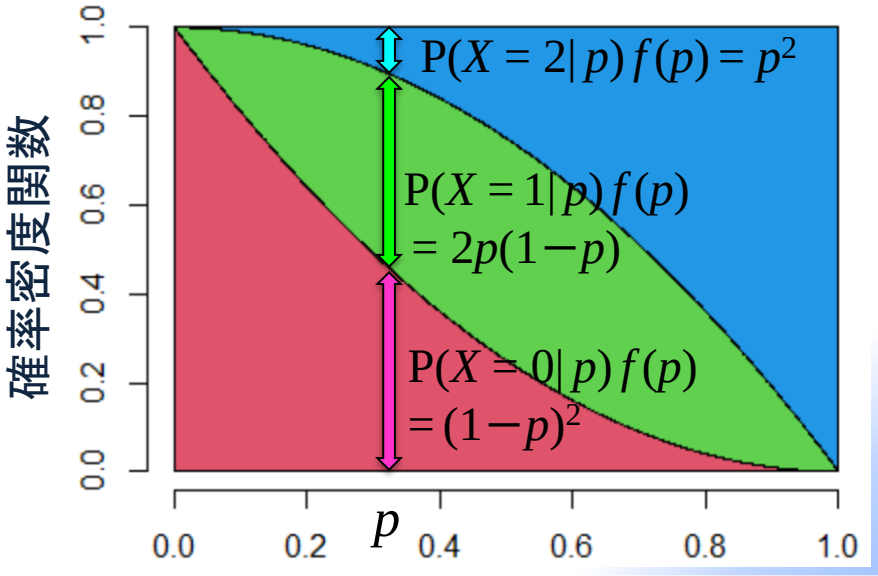
$$\mathrm{P}(X=x)=\int_0^1 \mathrm{P}(X=x\,|\,p)\,f(p)\,\mathrm{d}p=1/3\quad(x=0,1,2)$$

Step. 2で生成された X の値で色分け



p, X の生成数を無限にした極限

どの色の面積（$X=0,1,2$をとる確率）も1/3

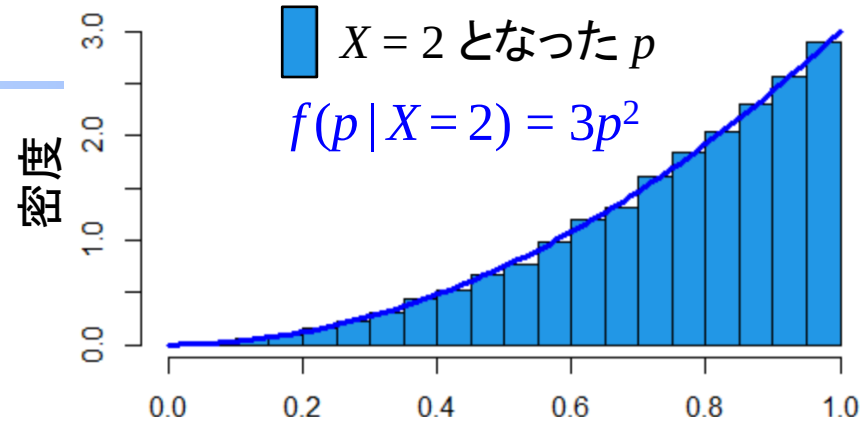


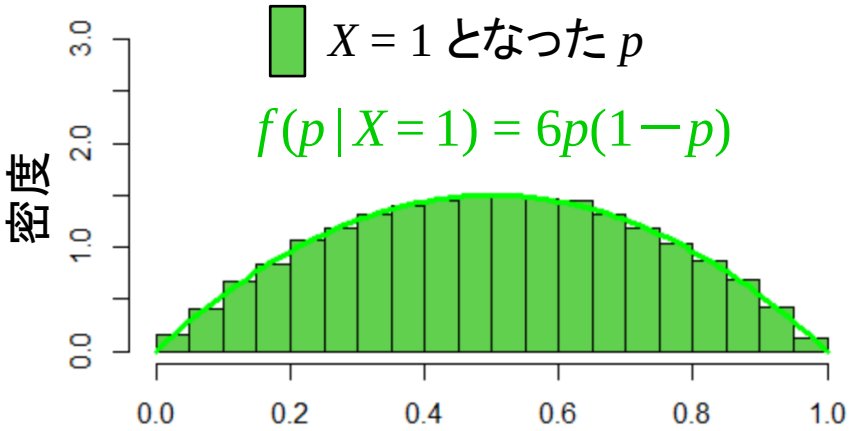
Xの値で分けたpの分布

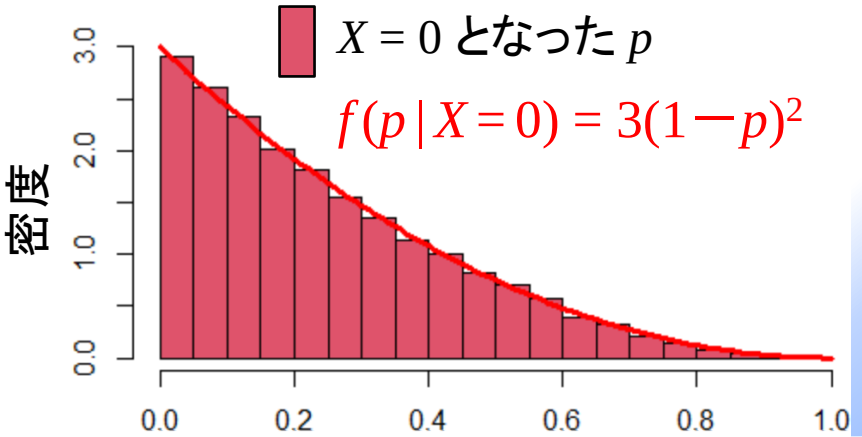
- Xの値ごとにpを分けてヒストグラムを描く
- 縦軸を密度 $=\dfrac{\text{相対度数}}{\text{階級幅}}$ として総面積 $=1$（全確率）にすると、それぞれ次の確率密度関数（事後密度）へと近づいていく

$$f(p \mid X=x)=\frac{\mathrm{P}(X=x \mid p) f(p)}{\mathrm{P}(X=x)}$$

$$=\begin{cases} 3p^2 & x=2\text{のとき} \\ 6p(1-p) & x=1\text{のとき} \\ 3(1-p)^2 & x=0\text{のとき} \end{cases}$$





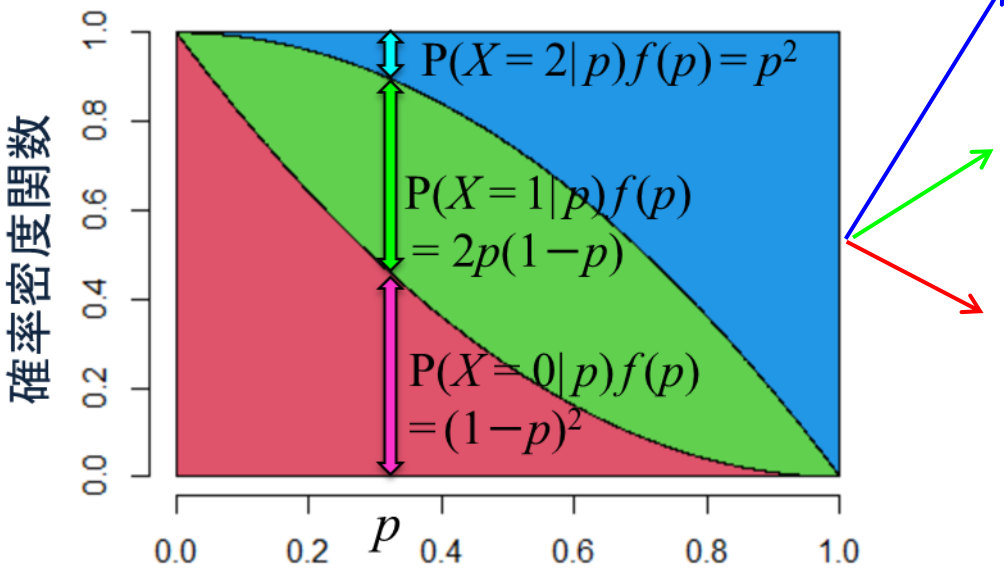


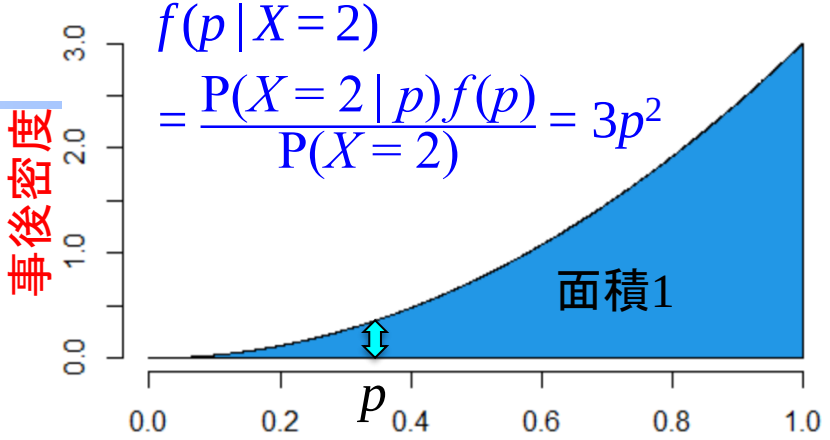
Xの値で分けたpの分布

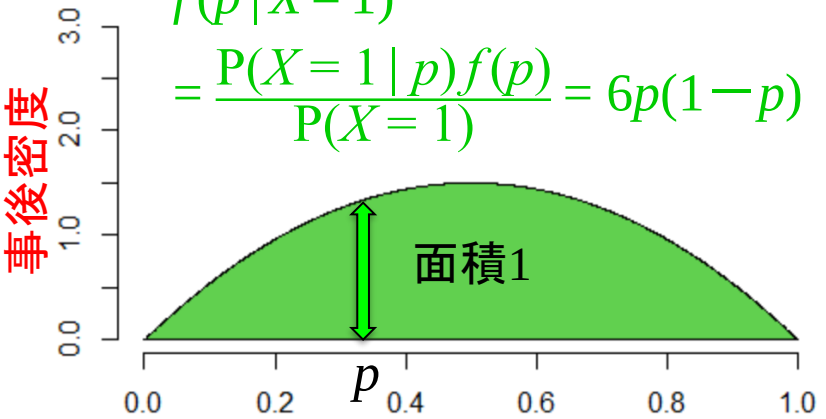
- 事後密度（前頁の密度の極限）は$\mathrm{P}(X=x\,|\,p)\,f(p)$に比例するが、総面積を1（全確率）にするため

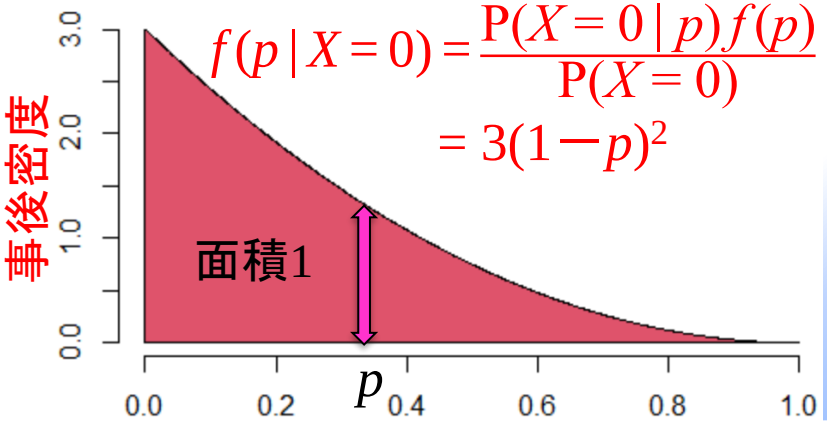
$$\mathrm{P}(X=x)=\int_0^1 \mathrm{P}(X=x\,|\,p)\,f(p)\,\mathrm{d}p$$

で割ることで前頁の式となる（ベイズの定理と呼ばれる）







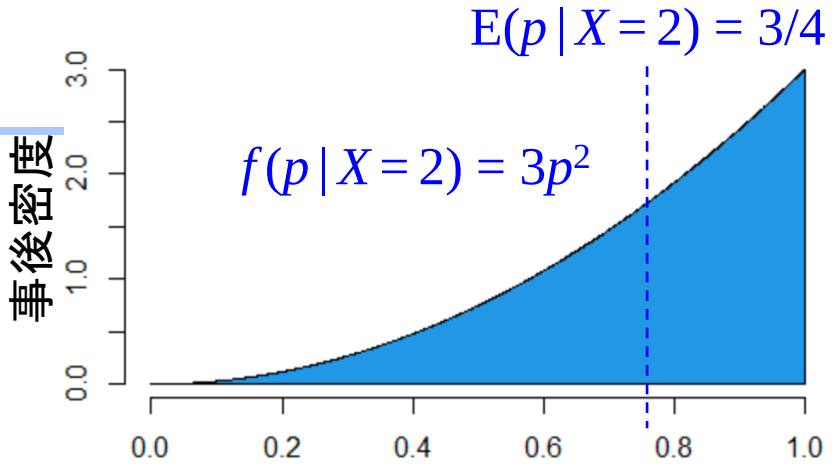


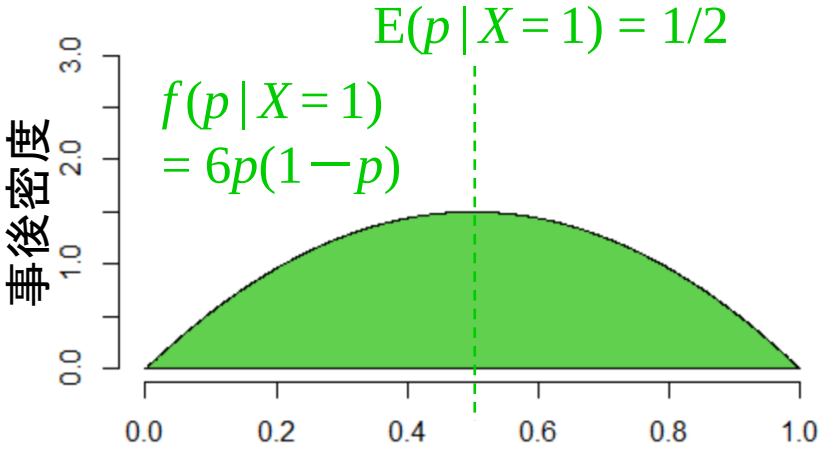
Xの値ごとのpの推定値

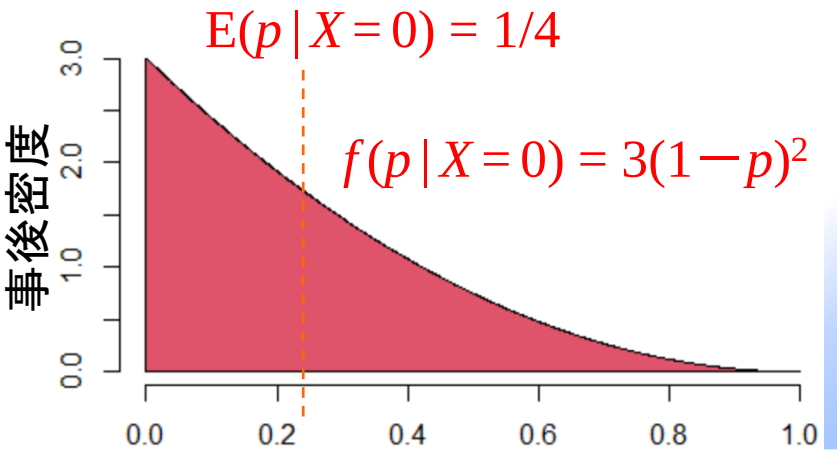
- 最良の推定値（Xの値ごとのpの平均）のシミュレート数を無限に増やしたときの極限は、**pの事後密度に関する平均**（pに対する二乗誤差の期待値を最小にする最良の推定値で**ベイズ推定値**という）となる

$$\mathrm{E}(p \mid X=x) = \int_0^1 p\, f(p \mid X=x)\,\mathrm{d}p$$

$$= \begin{cases} 3/4 & x=2\text{のとき} \\ 1/2 & x=1\text{のとき} \\ 1/4 & x=0\text{のとき} \end{cases}$$







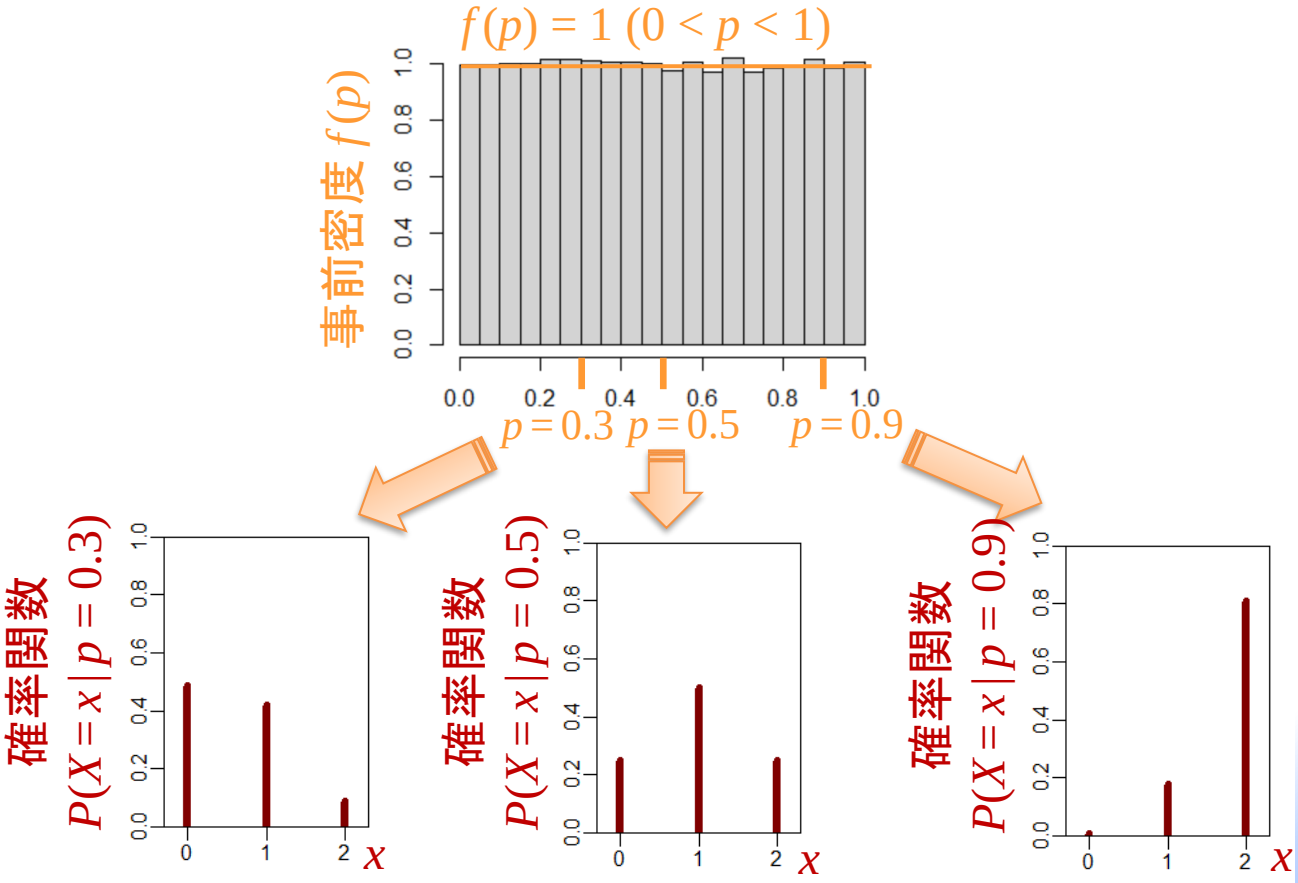
ベイズモデル

このゲームの p と X の階層構造はベイズモデルと呼ばれる:

- パラメータを生成する事前分布: $f(p) = 1 \quad (0 \leq p \leq 1)$
- データを生成する観測モデル: $\mathrm{P}(X = x \mid p) = {}_2\mathrm{C}_x p^x (1-p)^{2-x}$

事前分布による
パラメータ p
の生成

観測モデルによる
データ X
の生成



コイン投げシミュレーションゲームまとめ

- 次のゲームについて以下の問いに答えよ

Step. 1　0以上1未満の一様分布（事前分布）から乱数を生成して p とおく（確率密度関数 $f(p) = 1\ (0 \leq p \leq 1)$ ）

Step. 2 表が出る確率 p のコインを 2 回投げて表が出た数を X とおく（確率 $\mathrm{P}(X = x \mid p) = {}_2\mathrm{C}_x p^x (1-p)^{2-x}\ (x = 0,1,2)$ ）

問1：$X = x\ (x = 0,1,2)$ となる確率（周辺確率）は？

問2：$X = x\ (x = 0,1,2)$ となる p が従う分布（事後分布）は？

問3：$X = x\ (x = 0,1,2)$ となったときの p のベイズ推定値（平均二乗誤差を最小にする意味で最良の推定値）は？

コイン投げシミュレーションゲームのまとめ

問1：$X = x\ (x = 0,1,2)$ となる確率（周辺確率）は？

$$\mathrm{P}(X = x) = \int_0^1 \mathrm{P}(X = x \mid p) f(p) \mathrm{d}p = \frac{1}{3}$$

問2：$X = x\ (x = 0,1,2)$ となる p が従う分布（事後分布）は？

$$f(p \mid X = x) = \frac{\mathrm{P}(X = x \mid p) f(p)}{\mathrm{P}(X = x)} = \frac{3!}{x!(2-x)!} p^x (1-p)^{2-x}\ (0 \le p \le 1)$$

ベイズの定理　　ベータ分布

問3：$X = x\ (x = 0,1,2)$ となったときの p のベイズ推定値は？

$$\mathrm{E}(p \mid X = x) = \int_0^1 p f(p \mid X = x) \mathrm{d}p = \frac{x+1}{4}$$

コイン投げシミュレーションゲームのまとめ（一般化）

コイン投げの回数を n 回へと一般化すると次の結果となる

（$\mathrm{P}(X=x\,|\,p)={}_n\mathrm{C}_x\,p^x(1-p)^{n-x}\ (x=0,1,\ldots,n)$ ）

問1：$X=x\ (x=0,1,2)$ となる確率（周辺確率）は？

$$\mathrm{P}(X=x)=\int_0^1 \mathrm{P}(X=x\,|\,p)\,f(p)\mathrm{d}p=\frac{1}{n+1}$$

問2：$X=x\ (x=0,1,2)$ となる p が従う分布（事後分布）は？

$$f(p\,|\,X=x)=\frac{\mathrm{P}(X=x\,|\,p)f(p)}{\mathrm{P}(X=x)}=\frac{(n+1)!}{x!(n-x)!}p^x(1-p)^{n-x}$$

問3：$X=x\ (x=0,1,2)$ となったときの p のベイズ推定値は？

$$\mathrm{E}(p\,|\,X=x)=\int_0^1 p\,f(p\,|\,X=x)\mathrm{d}p=\frac{x+1}{n+2}$$

スポーツ選手の誕生月（予測）

- 誕生月の予測についても、各月の確率が合計1となる範囲で一様な分布を事前分布として仮定した場合の各月の確率の最良な推定値が次式で得られる

$$i\text{月が誕生月の確率の推定値} = \frac{i\text{月が誕生月の度数}+1}{\text{度数の合計}+12}$$

誕生月	4月	5月	6月	7月	8月	9月	10月	11月	12月	1月	2月	3月	計
度数	8	6	7	6	4	4	5	4	2	0	3	1	50
相対度数	0.16	0.12	0.14	0.12	0.08	0.08	0.10	0.08	0.04	0.00	0.06	0.02	1.00
度数＋1	9	7	8	7	5	5	6	5	3	1	4	2	62
確率の推定値	0.15	0.11	0.13	0.11	0.08	0.08	0.10	0.08	0.05	0.02	0.06	0.03	1.00

まとめ

- 統計学はデータを生成する確率的な構造の検証に興味が注がれるのに対して、機械学習はタスク（多くの場合は予測）の達成に主な焦点が当たる
- 統計学でも機械学習でもベイズモデルは重要な基礎であり、発展的かつ実用的なベイズモデルが多種多様に存在する
  - 時系列モデル（状態空間モデル）
  - 時空間モデル　→　佐野さんのご発表
  - 因果推論（グラフィカルモデル）
  - クラスタリング（テキストマイニング等）

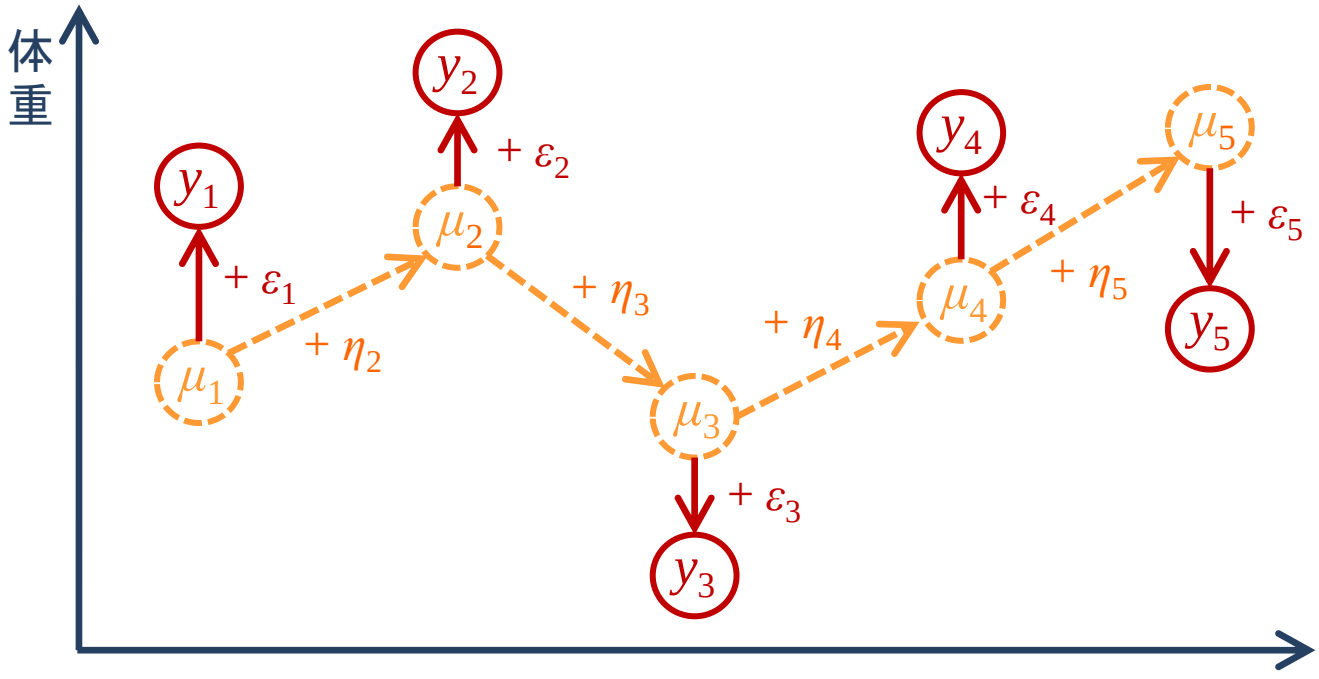
おまけ：ベイズモデルの応用例（状態空間モデル）

- 状態空間モデル：時系列データに対するベイズモデル

例）体重測定から測定誤差を除いた真の体重変化を推定

- 時点 t の真の体重：$\mu_t = \mu_{t-1} + \eta_t$ （η_t は前日からの増減）
- 時点 t の体重測定：$y_t = \mu_t + \varepsilon_t$ （ε_t は測定誤差）

y_t：データ（体重測定値）　μ_t：パラメータ（真の体重）



おまけ：ベイズモデルの応用例（状態空間モデル）

- 過去から現在までの真の体重 μ_t の変化を、データ（体重計測値）y_t に基づいて推定した
- 各時点の μ_t の推定値とその95%信頼区間が得られ、期間内の真の体重の増減を議論することができる