確率統計（２）

概要

1. 統計学の基礎
   - 1次元のデータ
   - 2次元のデータ
   - 相関
2. 確率の基礎
   - 確率変数と確率分布
   - 中心極限定理
   - 独立性
3. 仮説検定
   - t検定
   - カイ二乗検定
4. 推定
   - 点推定・区間推定
   - 最尤推定

補助資料：http://small-island.work/trial/
ユーザ名：trial
パスワード：trial

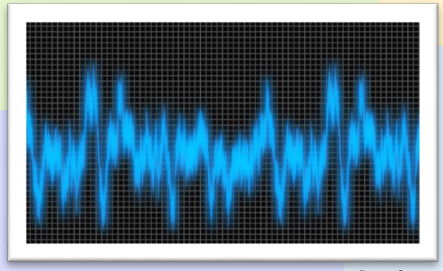
確率

代表値やデータのばらつき・不確実性を扱うための方法論で、確率を理解することで、ある事象が偶然起こったのかどうかをより深く理解することができる

世の中の多くのデータは確率的な現象を含んでいる

医療データ・記録データの不確実性

- 個人差
- 記録ミス
- 記録していない・できないデータ



センシングのノイズ

- 熱などによるセンサーの雑音
- 目的とする信号以外の信号

言語の曖昧さ

- 文法の揺れや表記の揺れ
- 意味の曖昧さ

自然科学・社会科学の中での確率的現象

- 人間や動物の行動
- 複雑すぎて確率的にしかわからない現象

確率変数（離散）

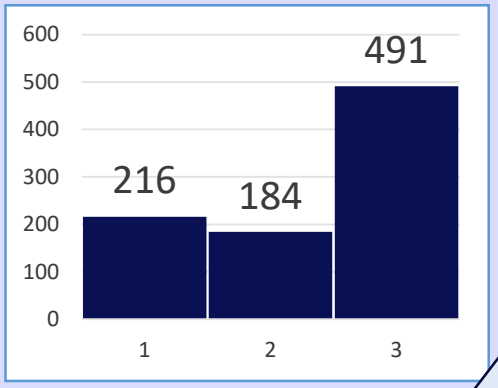
「確率変数」は、取りうる値の範囲である確率で値をとる変数のこと

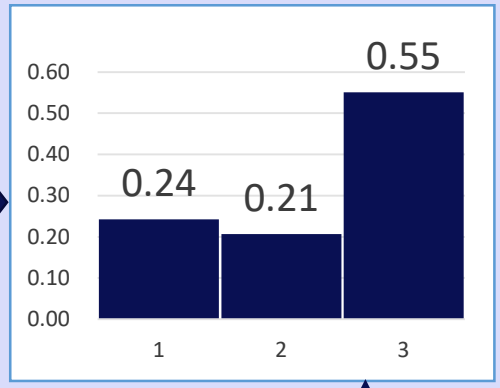
離散確率変数：取りうる値の範囲が離散的な確率変数

例：さいころを投げて出る目それぞれの目が出る確率は1/6であり、１～６の離散的な値をとることから、さいころを投げて出る目は離散確率変数であると言える

タイタニック号の乗客データ（https://www.kaggle.com/c/titanic/data, train）から乗客の階級１～３ (Pclass)（1の方が高い階級）についてヒストグラムを描く

Pclassに関するヒストグラム





全ての取りうる値を足すと１になるように正規化

ここでランダムに乗客を選んでその階級を見るとすると以下のように書ける

$P(\text{Pclass=1}) = 0.24$

$P(\text{Pclass=2}) = 0.21$

$P(\text{Pclass=3}) = 0.55$

Pclassは離散確率変数といえる

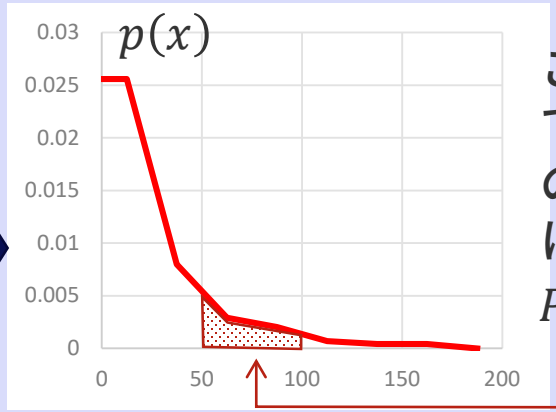
この時これをPclassの離散確率分布という

確率変数（連続）

連続確率変数：取りうる値の範囲が実数である確率変数

例：明日の気温は連続的な実数の中のいずれかの値をとることから確率変数であるといえる

タイタニック号の乗客データ（https://www.kaggle.com/c/titanic/data, train）から乗客の乗船料金 Fare（乗船料金は小数値を含む）についてヒストグラムを描く



ここでランダムに乗客を選んで例えば乗船料金が50~100の間にある確率を以下のように書く

$P(50<\text{Fare}<100) = \int_{50}^{100} p(x)dx$

この部分の面積

全ての定義域で積分すると１になるように正規化:$f(x)$

取りうる値の範囲が実数なのでFareは連続確率変数であり、$p(x)$を確率分布関数という

確率分布

確率変数の確率的なふるまいを表現する方法

大きく分けて、ノンパラメトリックな方法とパラメトリックな方法がある

ノンパラメトリック

分布の「かたち」を決めない方法
複雑な形を表現できるが、数式で解析的な計算には向いていない
例：ヒストグラムから決めた確率分布

パラメトリック

分布の「かたち」を数式で指定する方法
数式で表現できるため、確率を使った計算では便利
例：正規分布、二項分布、ポアソン分布など（次のスライド以降で解説）

利点
- 「かたち」が決まっているので比較的少ないデータでも分布を決められる
- 条件付き確率（後述）の計算など確率計算が比較的簡単にできる
- 確率モデル（機械学習講義にて）などでよく使われる

欠点
- 分布の形が現実と大きく異なるとうまく現実問題に適合しない

正規分布

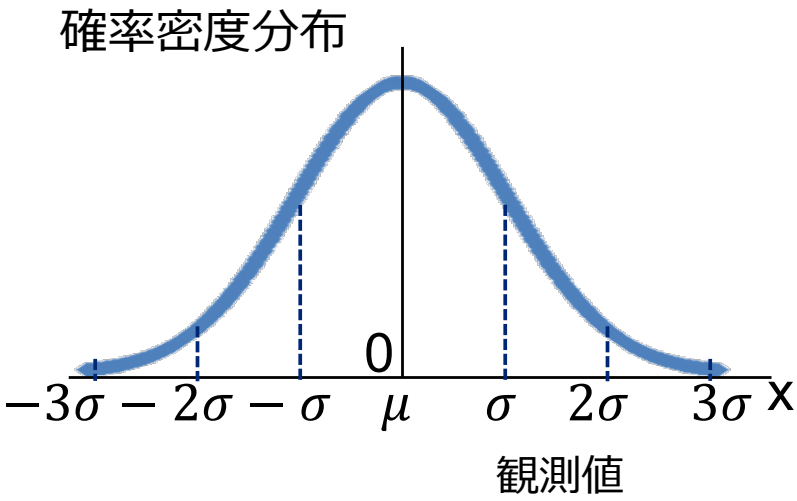
概要

- 各手法において、母集団分布が正規分布であると仮定することが多い
- 自然界や人間社会の多くの事象は、標本数が十分多ければ、正規分布に近づくものが多いためである
- ガウスが19世紀初頭に発見したことからガウス分布とも呼ばれる

数式

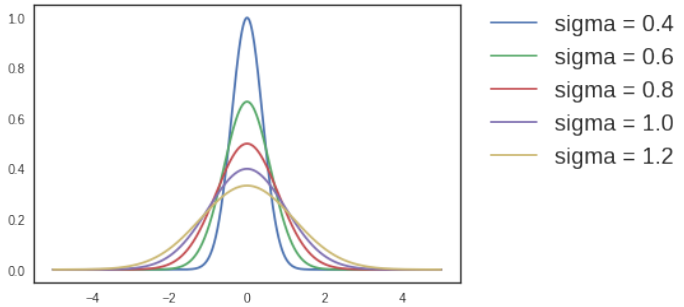
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- この分布の平均はμ（中心位置）、分散（ばらつき）はσ^2
- この正規分布を$N(\mu, \sigma^2)$と表記する
- 平均μに近いほど確率が高く、平均を中心に左右対称の釣鐘型の形状



性質

- xが$\mu \pm 1\sigma$、$\mu \pm 2\sigma$、$\mu \pm 3\sigma$に含まれる確率は、それぞれ68.3%、95.4%、99.7%となる
- 確率分布は平均と分散のみで決まる
- 2つの正規分布を四則演算しても正規分布



二項分布

概要

- 例えばコイントスで表が出た回数など二値をとる事象の発生回数が従う分布
- 発生回数なので離散的な確率変数に対応

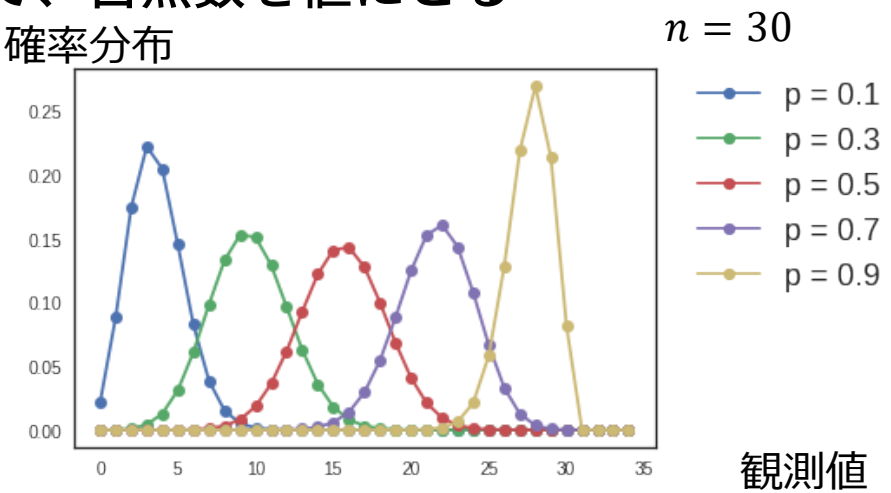
数式

$$P(X = k) = {}_nC_k \, p^k (1-p)^{n-k}$$

- この分布の平均はnp、分散（ばらつき）は$np(1-p)$
- 左右非対称の釣鐘型の形状をした分布となる
- 自然数のパラメータnに対して、自然数を値にとる

性質

- 二値ではなく多値の場合には多項分布に一般化できる



ポアソン分布

概要

- 事象の一定時間内での発生回数が従う分布
- 発生回数なので離散的な確率変数に対応

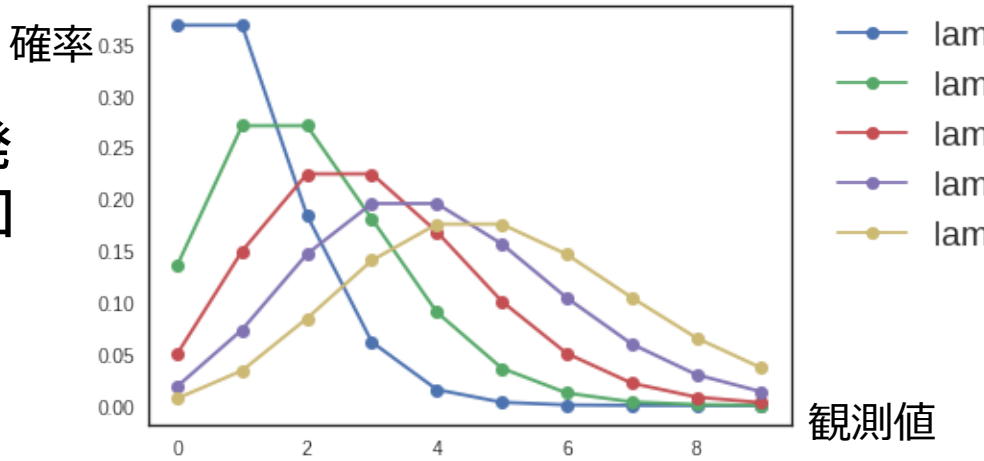
数式

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- この分布の平均はλ、分散はλ
- 左右非対称の釣鐘型の形状をした分布となる
- 正の実数のパラメータλに対して、自然数を値にとる

性質

- ある時間中に平均で λ 回発生する事象がちょうど k 回発生する確率



ガンマ分布

概要

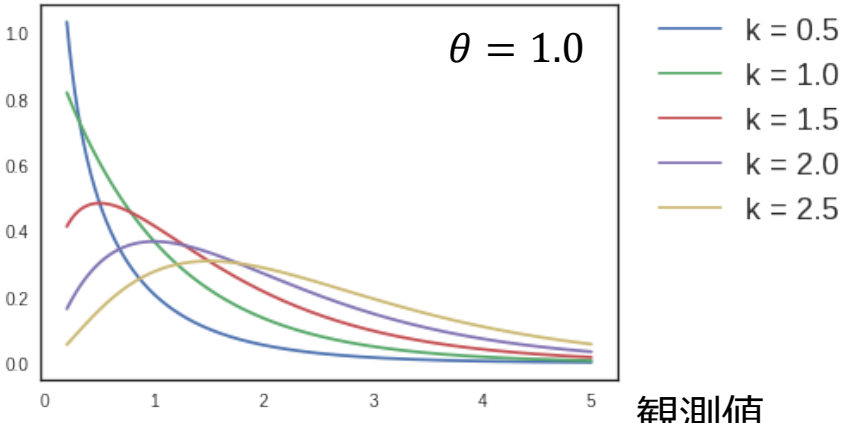
- 連続的な確率変数に対応
- 様々な分布の一般形になっている
  例えば部品の寿命など切り替わりまでの時間が従う指数分布

数式

$$p(x) = \frac{1}{\Gamma(k)\theta^k}\, x^k e^{-x/\theta}$$

- この分布の平均は$k\theta$、分散は$k\theta^2$
- パラメータk, θ に対して、正の実数値を値にとる

確率密度分布



性質

特に $k = 1$ である場合、このガンマ分布は尺度母数（平均値）を θとする指数分布と呼ばれる

$k = n/2$ $(n = 1, 2, \ldots)$ かつ $\theta = 2$ である場合、ガンマ分布は自由度 n のカイ二乗分布と呼ばれる
（第3回確率統計：e-learning の「検定」で利用される分布）

確率変数の性質: 確率変数に関数を適用したものは確率変数

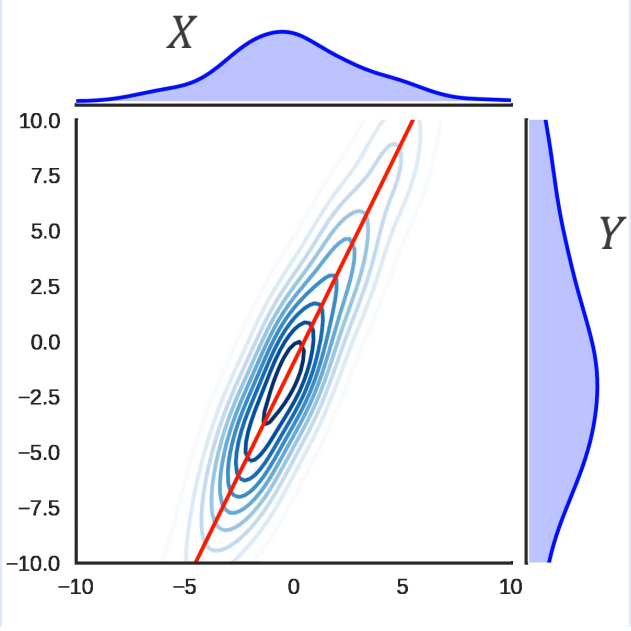
ここまでに述べた直接確率変数の分布を定義するほかに、すでにある確率変数から新しく確率変数を作ることもできる

確率変数に関数を適用したものは確率変数なので、
確率変数を定数倍したものや定数を足したものも確率変数

例１：Xが確率変数ならば、

$$Y = 2X - 2$$

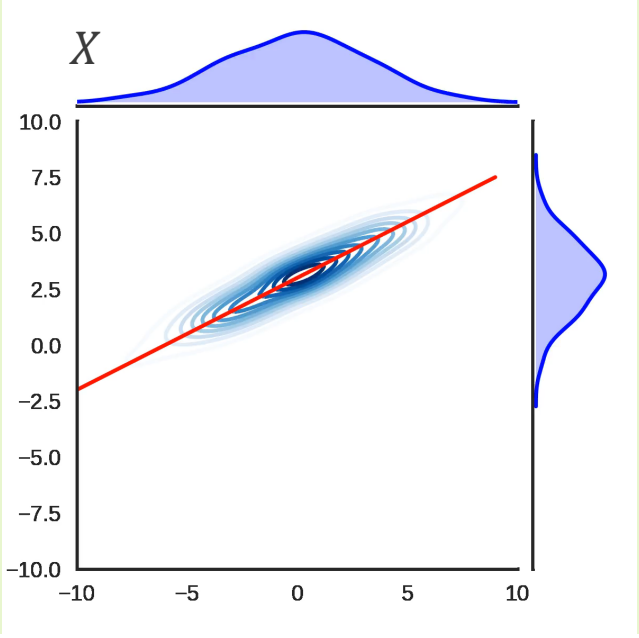
で変換したYも確率変数



例２：Xが確率変数ならば、

$$Y = \mathrm{X}/2 + 2$$

で変換したYも確率変数

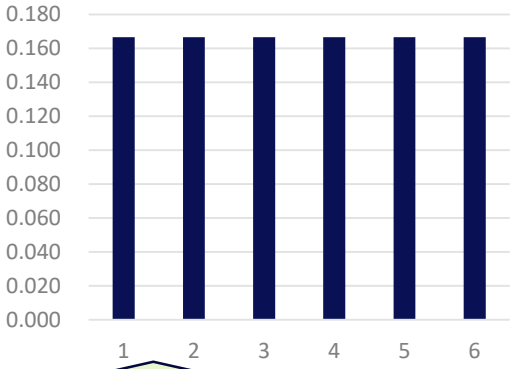


中心極限定理

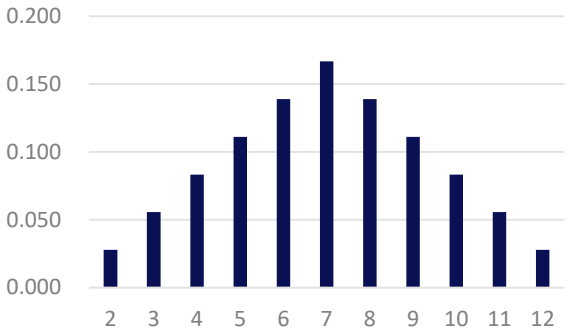
（同一で独立な）確率変数を多く足し合わせることでできる新しい確率変数は
正規分布に従う

例：さいころをn回も投げて、それぞれ出る目を表す確率変数を $X_1, X_2, \ldots, X_n$ ▨する

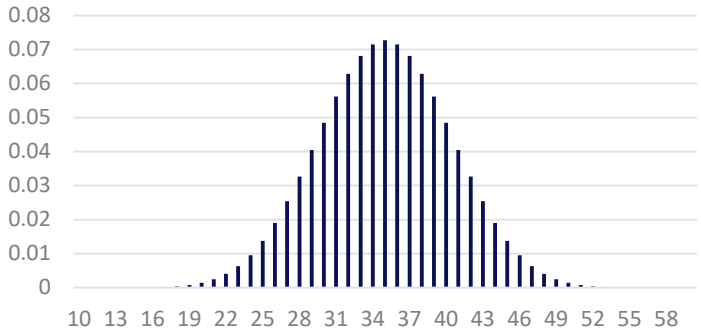
X_1の確率分布



$X_1 + X_2$の確率分布



$X_1 + X_2 + \cdots + X_{10}$の確率分布



このような分布を一様分布と呼ぶ

確率変数を多く足し合わせることで、正規分布に近い形になっていくことがわかる
一様分布以外の歪なサイコロを用いても、同様のことが起こる

この定理は統計の至る所で重要な役割を果たす
（第4回確率統計：e-learningの「推定」でも重要な役割を果たす）

複数の確率変数間の関係を議論するための様々な確率

多変量のデータや二つのデータの関係性を分析するには、複数の確率変数間の関係を議論する必要がある。

同時確率(分布)：二つ以上の事象が同時に起こる確率

例：サイコロを二回投げて１回目で１、２回目で６が出る確率

$$P(X_1 = 1, X_2 = 6) = \frac{1}{36}$$

X_1:１回目に出た目を表す確率変数
X_2:２回目に出た目を表す確率変数

周辺確率(分布)：二つ以上の事象があるときに一つの事象のみに注目した確率

同時分布から計算できる

$$P(Y = y) = \sum_x P(X = x, Y = y)$$

連続の場合:

$$p(y) = \int_x p(x, y)dx$$

例：サイコロを二回投げて２回目で６が出る確率

$$P(X_2 = 6) = \sum_{x=1}^{6} P(X_1 = x, X_2 = 6) = \frac{1}{6}$$

条件付き確率(分布)：ある事象 y が起こった条件下での別の事象 x の起こる確率

同時分布と周辺分布から計算できる

$$P(X = x|Y = y) = P(X = x, Y = y)/P(Y = y)$$

連続の場合:

$$p(x|y) = p(x, y)/p(x)$$

複数の確率変数間の関係を議論するための様々な確率（例）

問題：サイコロを二回投げて小さい方の数字を表す確率変数Y_1と大きい方の数字Y_2の確率を考える

同時確率：以下の表にまとめられる（同じ数字のみ1/36、残りは1/18）

$Y_1 \backslash Y_2$	1	2	3	4	5	6
1	1/36	1/18	1/18	1/18	1/18	1/18
2	-	1/36	1/18	1/18	1/18	1/18
3	-	-	1/36	1/18	1/18	1/18
4	-	-	-	1/36	1/18	1/18
5	-	-	-	-	1/36	1/18
6	-	-	-	-	-	1/36

$P(Y_1 = 1) = 11/36$
$P(Y_1 = 2) = 9/36$
$P(Y_1 = 3) = 7/36$
$P(Y_1 = 4) = 5/36$
$P(Y_1 = 5) = 3/36$
$P(Y_1 = 6) = 1/36$

周辺確率：Y_1のみに注目した確率$P(Y_1 = y)$は横方向の総和

条件付確率：例えば$Y_1 = 3$が与えられたときの$Y_2 = 4$の確率

$$P(Y_2 = 4 \mid Y_1 = 3) = \frac{P(Y_2 = 4, Y_1 = 3)}{P(Y_1 = 3)} = \frac{1/18}{7/36} = \frac{2}{7}$$

複数の確率変数間の関係を議論する

多変量のデータや二つのデータの関係性を見たい時には、複数の確率変数間の関係を議論する必要がある

独立: 二つの確率変数の同時確率がそれぞれの確率の積で表すことができる

$$P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)\,P(X_2 = x_2)$$

例1：サイコロを二回投げて、1回目と２回目の目はこの式を満たすので独立である

例2：一つ前のスライドのサイコロを二回投げて小さい方の数字を表す確率変数Y_1と大きい方の数字Y_2の確率を考える。この時$P(Y_1 = y_1, Y_2 = y_2) \neq P(Y_1 =$

無相関（相関係数＝０）

$$E[(x_1 - \mu_1)(x_2 - \mu_2)] = 0$$
$$\mu_1 = E[x_1], \qquad \mu_2 = E[x_2]$$

（ここでは、第1回確率統計：e-learning にて解説した相関係数の分子と同等の式を確率を用いて記述している）

$E[\ \ ]$は期待値（重み付き平均）と呼ばれ以下で計算される

$$E[x] = \sum_x x\,P(X = x)$$

右辺の確率変数Xは左辺でよく省略されるので注意

条件付き独立：二つの確率変数がある変数の条件付き確率について独立である

$$P(X_1 = x_1, X_2 = x_2 \mid Z = z) = P(X_1 = x_1 \mid Z = z)\,P(X_2 = x_2 \mid Z = z)$$

参考資料等

ここでは確率や確率変数をしっかりと定義せずに、実用上ある程度問題ないレベルでその性質について記述している

確率のきちんとした定義は「確率論」の教科書を参照すること

本e-learningでは確率変数の扱いに慣れることに重点を置き、実用上重要だが、発展的な内容については「機械学習」等の他の講義にて解説する

- ベイズの定理を含む、ベイズ統計へとつながる項目
- ベイジアンネットワーク等の確率モデルや統計的機械学習へとつながる項目