広島大学

データサイエンス基礎

Fundamental Data Science

第10回 確率変数と確率分布

広島大学　数理統計学グループ

どのくじを引く?

くじ引き1

	A賞	B賞	C賞
賞金 (円)	1000	500	100
本数 (本)	1	3	6

くじ引き2

	A賞	B賞	C賞
賞金 (円)	1000	500	0
本数 (本)	2	2	6

くじ引き3

	A賞	B賞	C賞
賞金 (円)	1000	400	100
本数 (本)	2	1	7

どのくじを引こう??

私たちがナビゲートします!

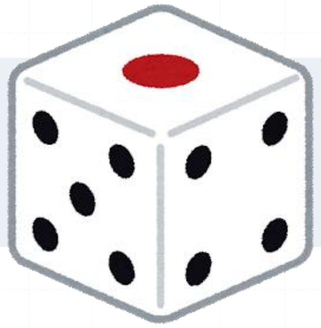
今日の内容

- ▶ **確率変数**
- ▶ **確率分布**
- ▶ **確率変数の期待値, 分散**

確率変数と確率分布

用語の復習

サイコロを1回投げたとき，**1の目が出た**



事象や標本空間は言葉でも数字でかいても OK

- **試行**
  偶然性を伴う実験や調査
- **事象**
  試行の結果としてありうる事柄で，標本空間の一部分　{1の目が出る} とか {1}
- **標本空間（全事象）Ω**
  試行の際に起こりうる結果の全体
  {1の目が出る, 2の目が出る, 3の目が出る, 4の目が出る, 5の目が出る, 6の目が出る} とか {1,2,3,4,5,6}

確率変数

でも言葉より数字のほうが数学的に扱いやすい

もし言葉でかいたなら… 言葉でかいた事象を**数字表記に置き換える**!!

確率変数

事象を数字に置き換える変換器のこと

例

サイコロ

1の目が出る ⇒ **1**
2の目が出る ⇒ **2**
と変換

数字をそのまま使用

例

試合の勝敗

勝ち ⇒ **1**
負け ⇒ **−1**
と変換

数字は区別のためだけに使用

問い

2枚のコインを同時に投げる試行を考える.

このとき, 標本空間を $\Omega = \{(表,表), (表,裏), (裏,表), (裏,裏)\}$ とかく.

このとき, 3つの確率変数 X, Y, Z を考えた

確率変数 X

	(表, 表)	(表, 裏)	(裏, 表)	(裏, 裏)
X	4	3	2	1

確率変数 Y

	(表, 表)	(表, 裏)	(裏, 表)	(裏, 裏)
Y	1	0	0	-1

確率変数 Z

	(表, 表)	(表, 裏)	(裏, 表)	(裏, 裏)
Z	2	1	1	0

Q いま, 試行の結果として **表が出る枚数**に興味の対象があるとき, どの確率変数が最も適切と考えられるか?

興味の対象をそのまま確率変数に反映させれるとよい!!

確率変数に対する確率

例 **2枚のコインを同時に投げる**

確率変数 X ： 表がでる枚数

確率変数に対する確率

$X=0$ である確率: $P(X=0)=\dfrac{1}{4}$ 表が0枚の確率

$X=1$ である確率: $P(X=1)=\dfrac{2}{4}$ 表が1枚の確率

$X=2$ である確率: $P(X=2)=\dfrac{1}{4}$ 表が2枚の確率

確率変数 X	0	1	2	計
確率	1/4	2/4	1/4	1

この表のことを
確率分布と呼びます

確率分布

一般的にかくと…

確率変数 X	x_1	x_2	$\cdots$	x_n	計
$P(X=x_k)$	p_1	p_2	$\cdots$	p_n	1

$$P(X = x_k) = p_k \; (k = 1, \ldots, n)$$

確率変数のとる値とその確率の対応関係を**確率分布**（または単に**分布**）といい，
確率変数はその確率分布に従うという

※ 確率変数に対する全ての確率 ＝ 確率分布 と思えば OK

確率分布は現象を理解するための道具

確率分布がわかれば，
その現象が起こる
確率的な仕組みがわかる!!

練習問題 (1)

Q A賞, B賞, C賞のどれかが当たる10本のくじがある.
10本からくじを1本引くときの賞金額を X 円とする.

くじ引き 1

	A賞	B賞	C賞
賞金 (円)	1000	500	100
本数 (本)	1	3	6

このとき, 確率変数 X が従う確率分布を求めよ.
(つまり, 下の表を埋めよ)

確率変数 X	1000	500	100	計
確率				1

離散型と連続型
（もっと詳しく!!）

確率変数の分類

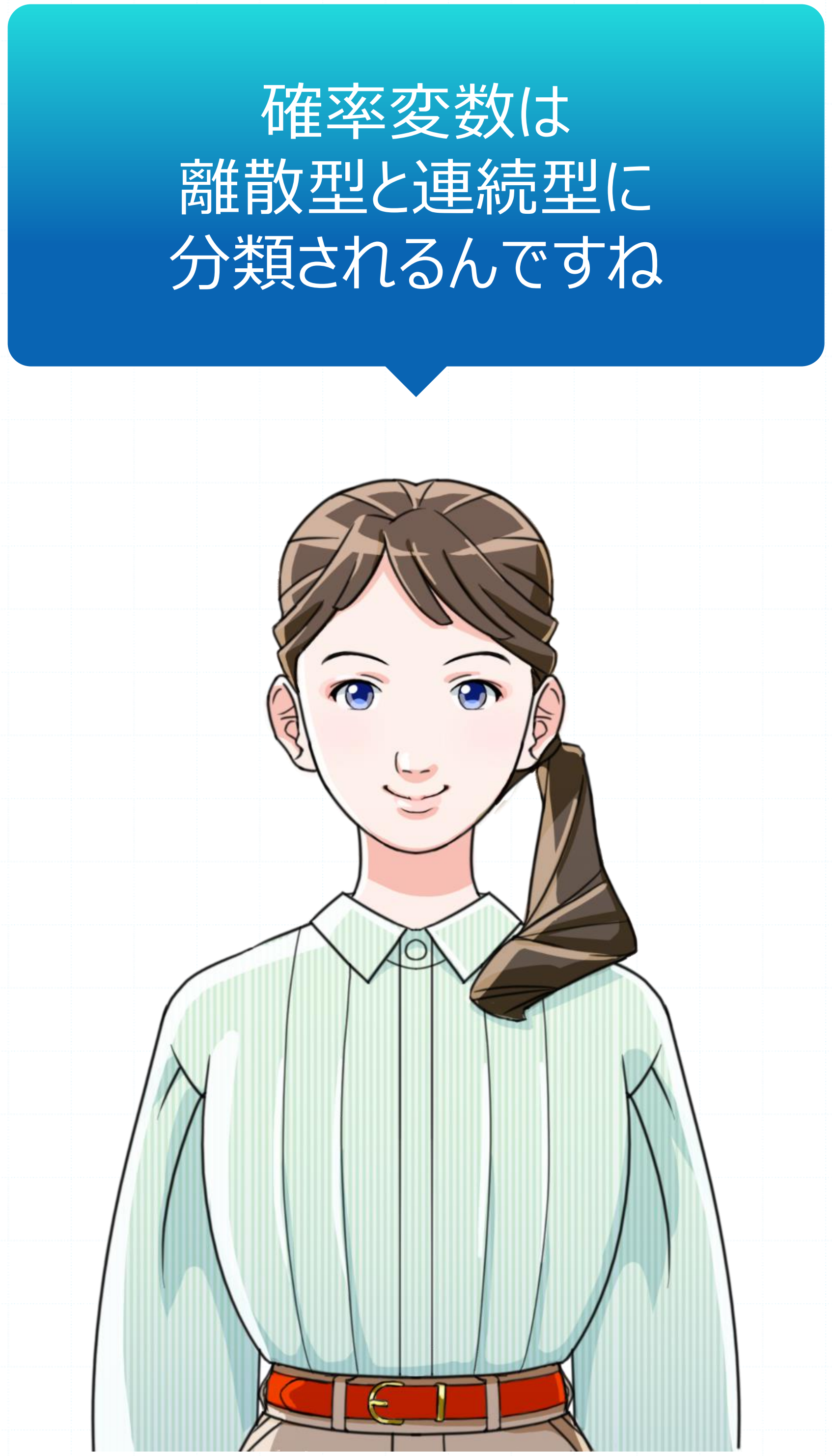
離散型確率変数

確率変数の値が**とびとびの値（離散値）**をとるとき，
離散型確率変数という

連続型確率変数

確率変数の値が**連続的な値**をとるとき，
連続型確率変数という

確率の扱い方が離散型，連続型でかわる!!



離散型確率変数の例

ここまでで扱った確率変数はすべて離散型!!

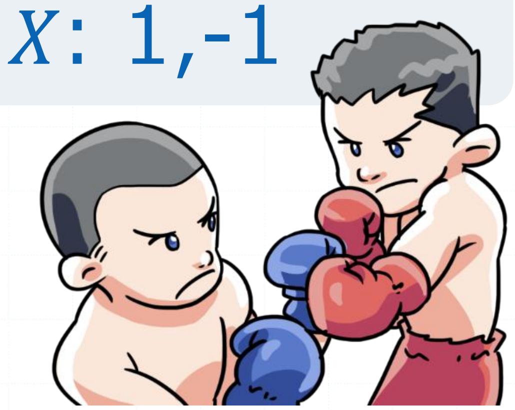
サイコロを投げて出る目の数

確率変数 W: 1,2,3,4,5,6

とびとびの値

試合の勝ち, 負け

確率変数 X: 1,-1

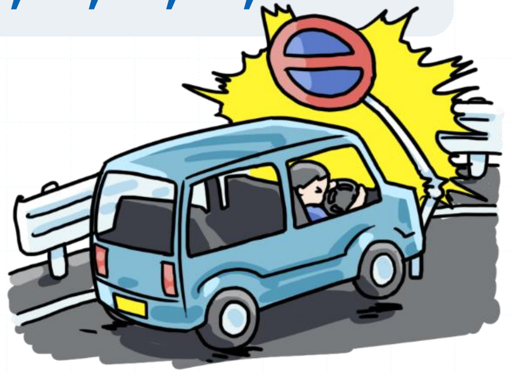


くじを引いたときの賞金

確率変数 Y: 1000,500,100

1年間で起こる交通事故の件数

確率変数 Z: 0,1,2,3,...



離散型分布

離散型分布

離散型確率変数 X が従う分布を離散型分布という．

X	x_1	x_2	$\cdots$	計
$P(X = x_k)$	p_1	p_2	$\cdots$	1

$$P(X = x_k) = p_k \ (k = 1,2, \dots)$$

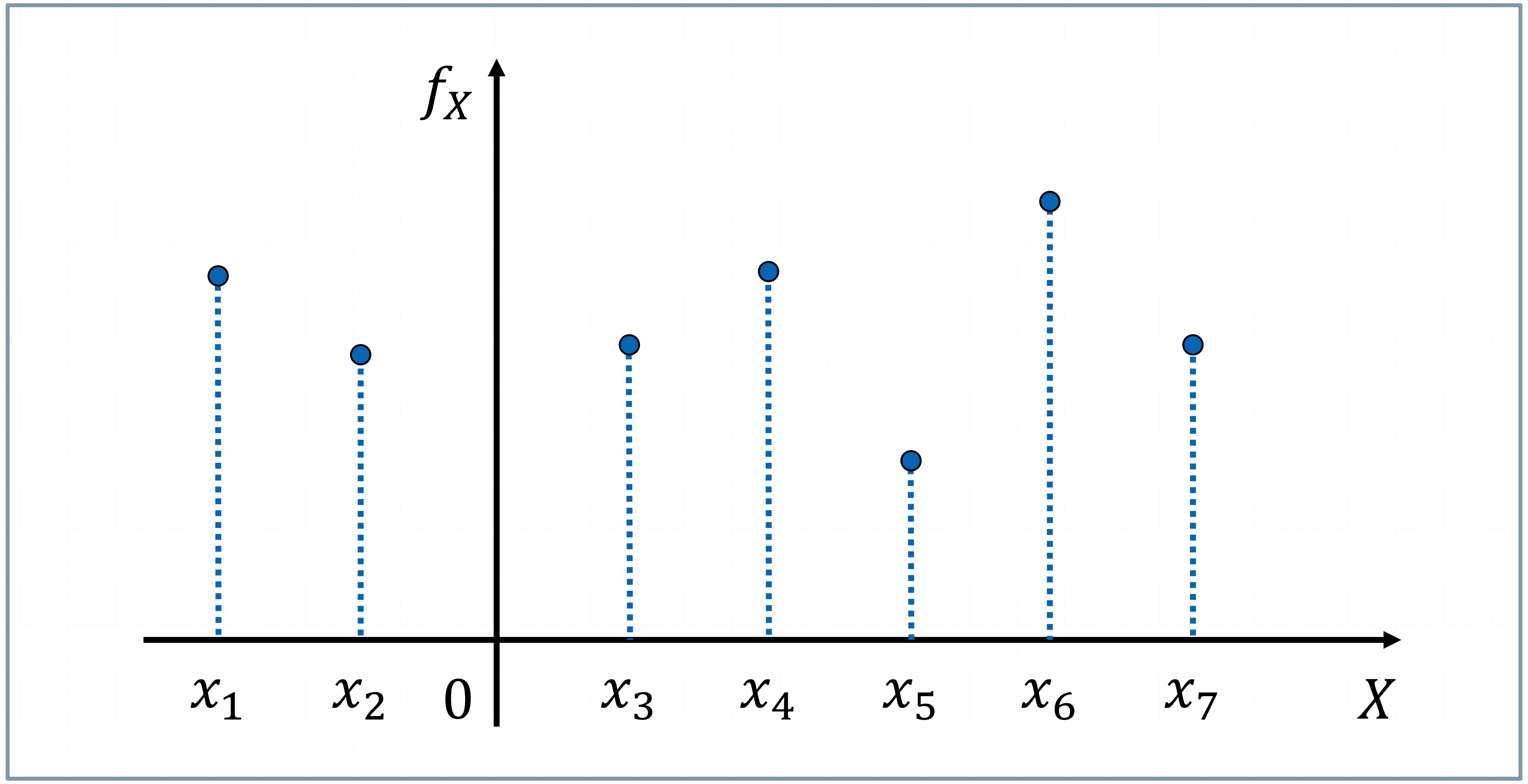
このとき，$P(X = x_k)$ を X の**確率関数**といい，$f_X(x_k)$ または f_X と表す．

確率関数 f_X は**確率変数1つ1つの値の確率**そのもの

f_X をすべて足すと 1: $f_X(x_1) + f_X(x_2) + \cdots = 1$

確率関数

例 **確率関数 $f_X(x_k) = P(X = x_k)$ のグラフ**



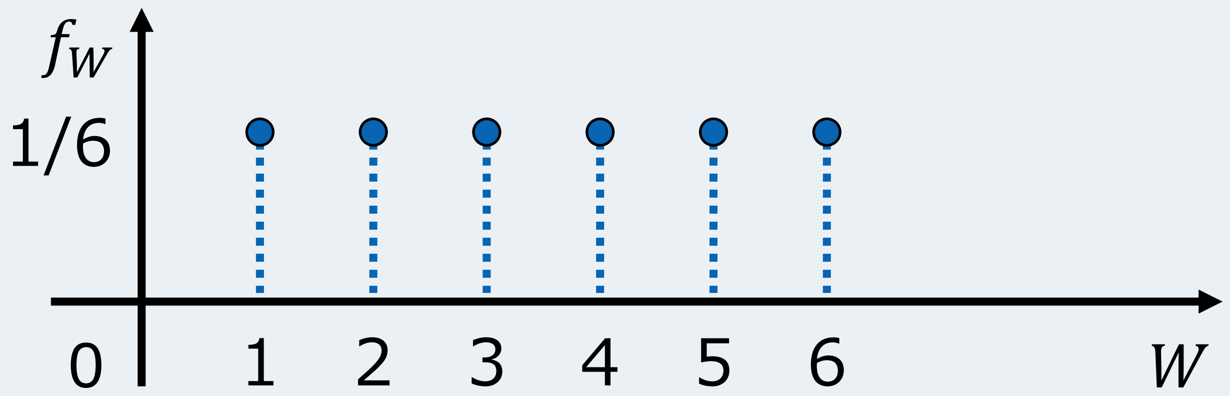
確率変数が離散型なので
横軸の値がとびとびに
なっているのがわかりますね

縦軸の値が確率 $P(X = x_k)$ を表す

確率関数の例

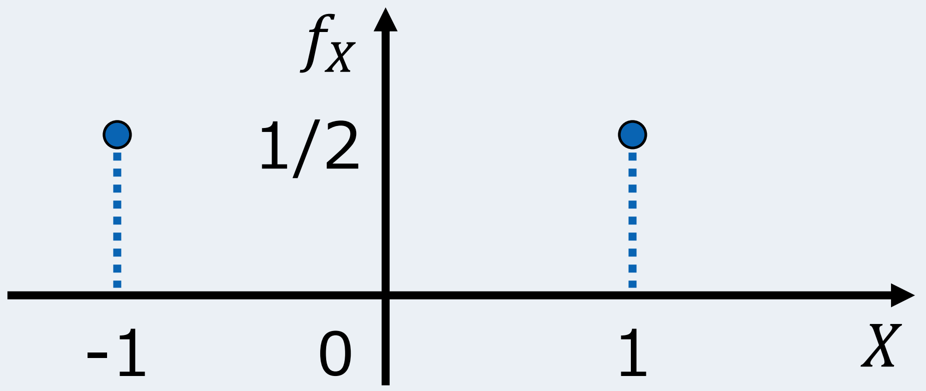
サイコロを投げて出る目の数

確率変数 W: 1,2,3,4,5,6



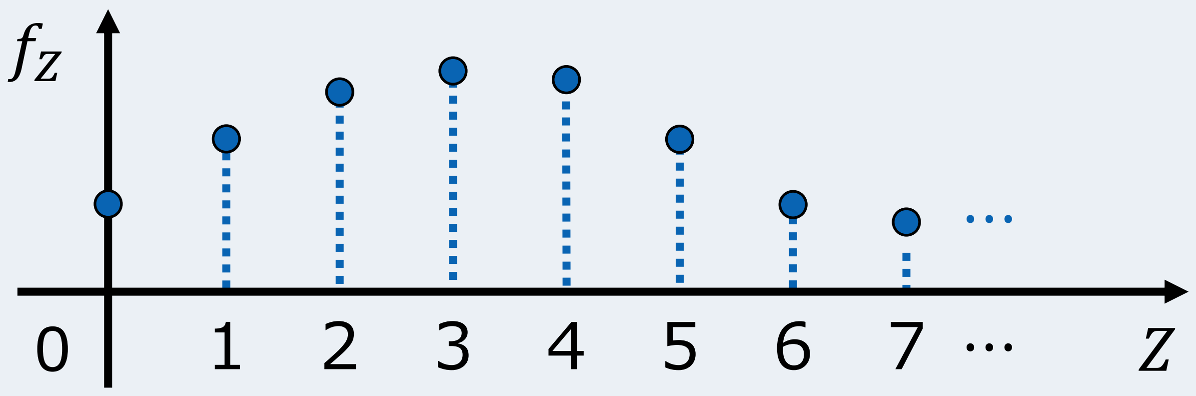
試合の勝ち, 負け

確率変数 X: 1,-1



1年間で起こる交通事故の件数

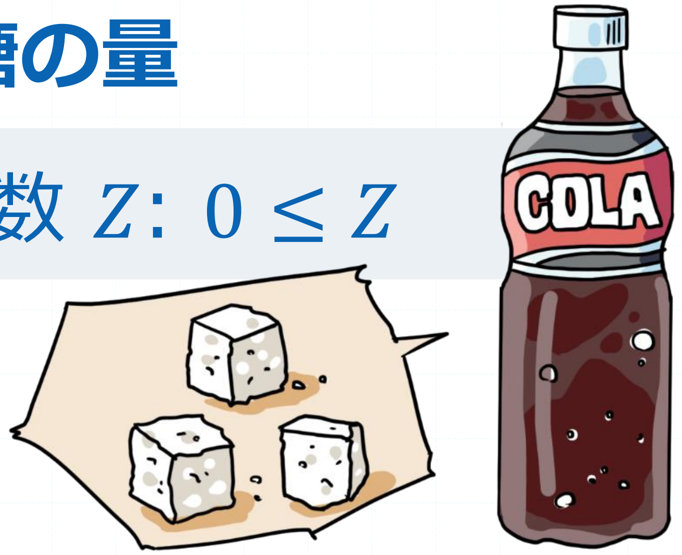
確率変数 Z: 0,1,2,3,…



連続型確率変数の例

自販機で買ったコーラに含まれる砂糖の量

確率変数 Z: $0 \leq Z$

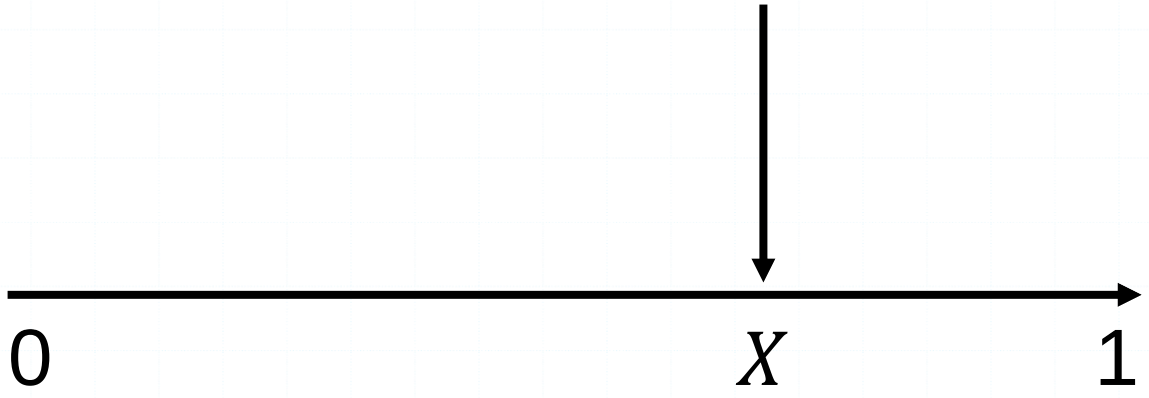


身長

確率変数 Y: $0 \leq Y$

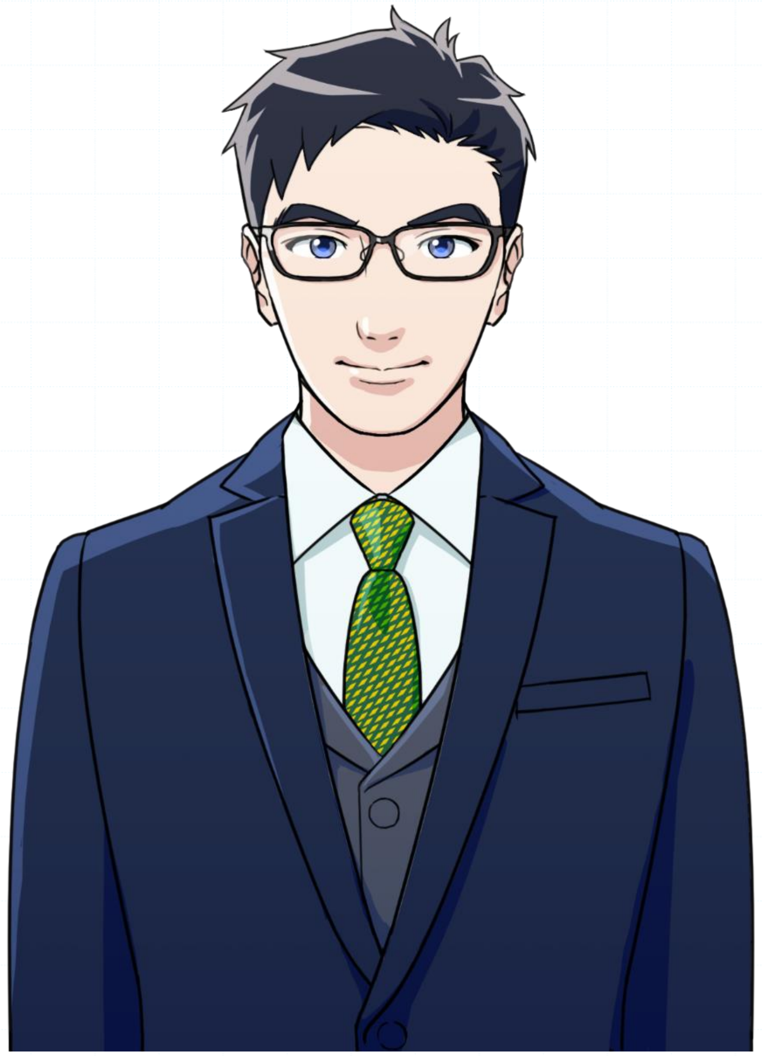
0から1の数直線上にランダムに針を落としたときの値

確率変数 X: $0 \leq X \leq 1$



連続的な値

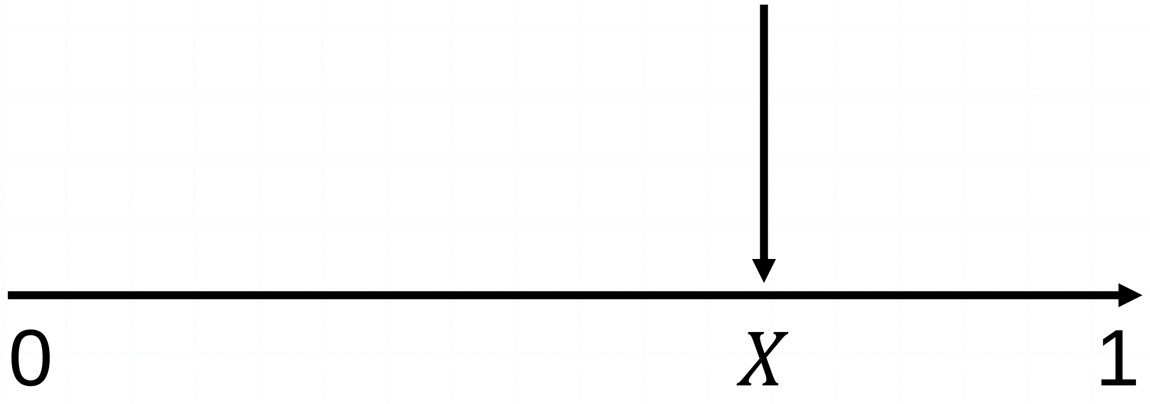
では
連続型確率変数は
どんなものでしょう



連続型確率変数の確率

0から1の数直線上に
ランダムに針を落としたときの値

確率変数 X: $0 \leq X \leq 1$



無限に多くの数から
0.5という数だけを当てるのは
無理だから確率が0なんですね

Q X が0.5となる確率 $P(X = 0.5)$ は??

A $\boldsymbol{P(X = 0.5) = 0}$

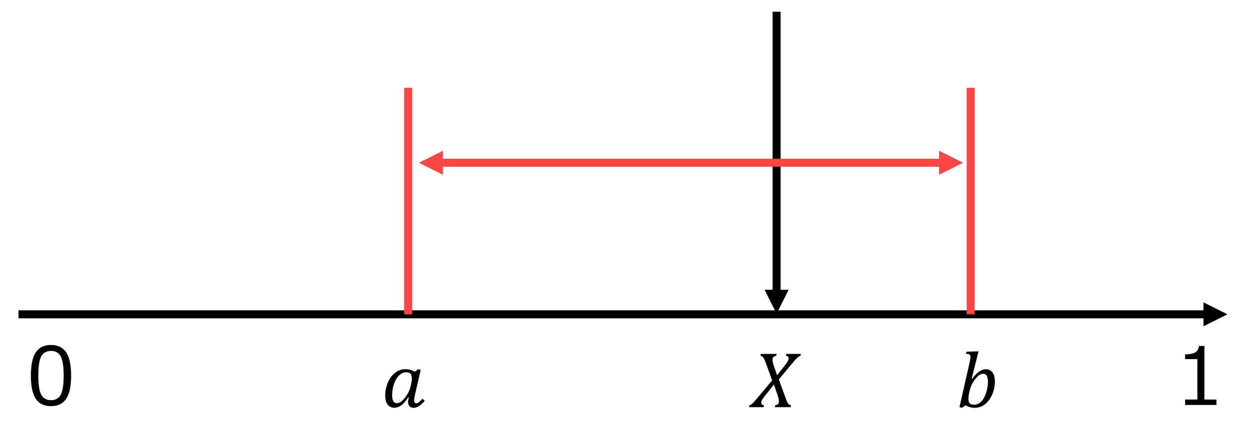
0と1の間には無限に多くの数がある
0.1,0.01,0.001,…

連続型確率変数は
1つの値をとる確率を計算しても0

連続型確率変数の確率

1つの値でなく，**幅をもたせた (範囲の) 確率**で考える

X が a 以上 b 以下である確率 $P(a \leq X \leq b)$



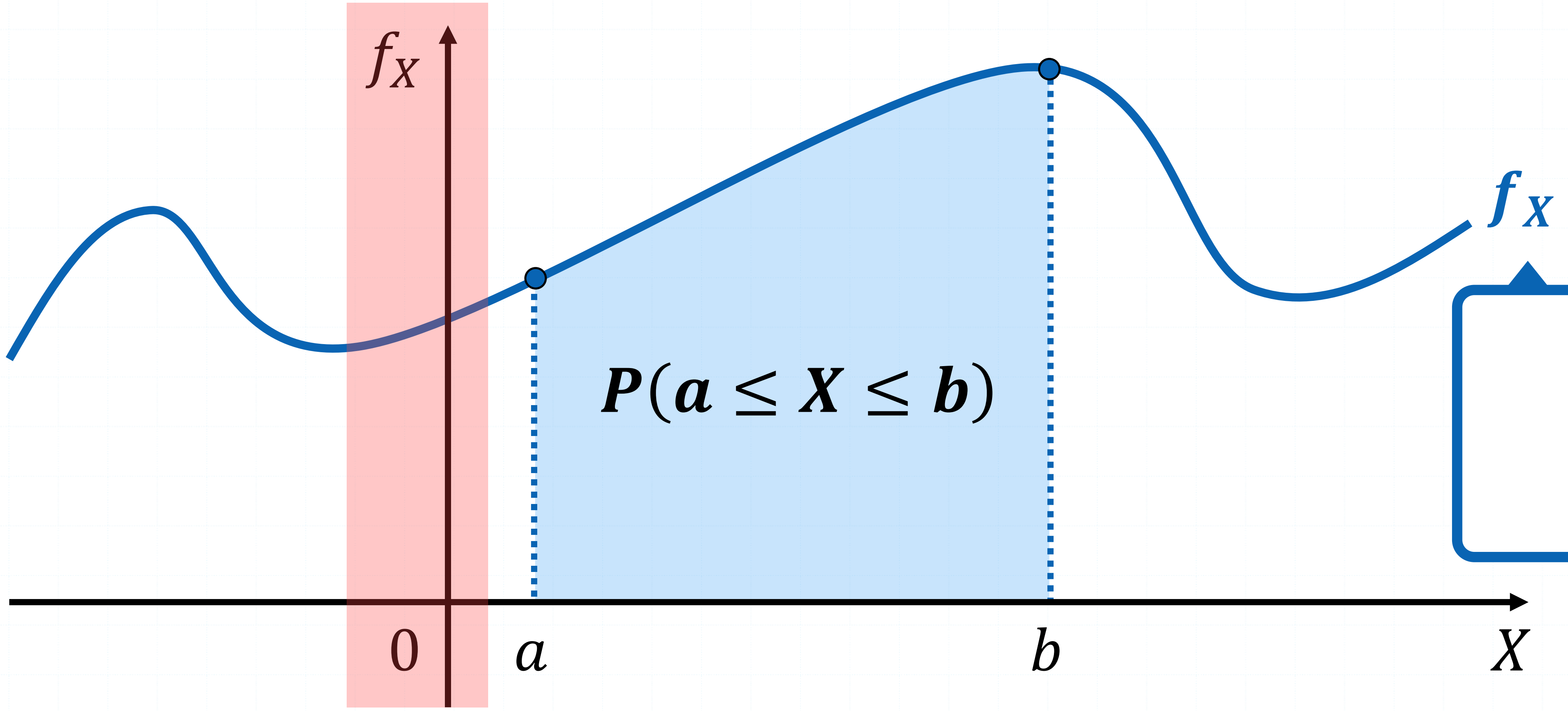
$$P(a \leq X \leq b) = \frac{\text{区間の幅}}{\text{全体の区間の幅}} = \frac{b-a}{1-0} = b-a$$

例 $P(0 \leq X \leq 0.5) = 1/2$

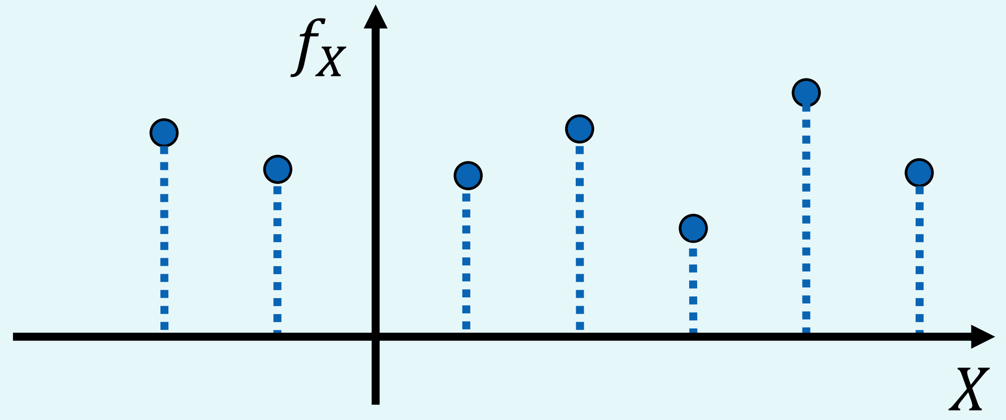
連続型確率変数は幅をもたせた確率を扱う

連続型確率変数の確率

確率を視覚的に表す
幅をもつ確率を**グラフの面積**で表す!!



離散分布: 縦軸が確率



f_X は面積に合う
ある曲線

「 f_X と $X=a,b$ で囲まれた面積」=「幅の確率」

連続型分布

連続型分布

連続型確率変数 X のとる幅の確率が，$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx$

と表されるとき，f_X を X の**確率密度関数**といい，
この X の分布を連続型分布という

$\int$ は積分記号
積分 ＝ 面積 と思えば OK

$P(a \leq X \leq b)$ は f_X と $X = a, b$ で囲まれた**面積**で表す
※ グラフの縦軸の値 (f_Xの値) は確率ではないことに注意!!

1つの値をとる確率：$P(X = a) = P(a \leq X \leq a) = \int_a^a f_X(x)\,dx = 0$

f_X を与えれば $P(a \leq X \leq b)$ が定まる．その逆も言える．

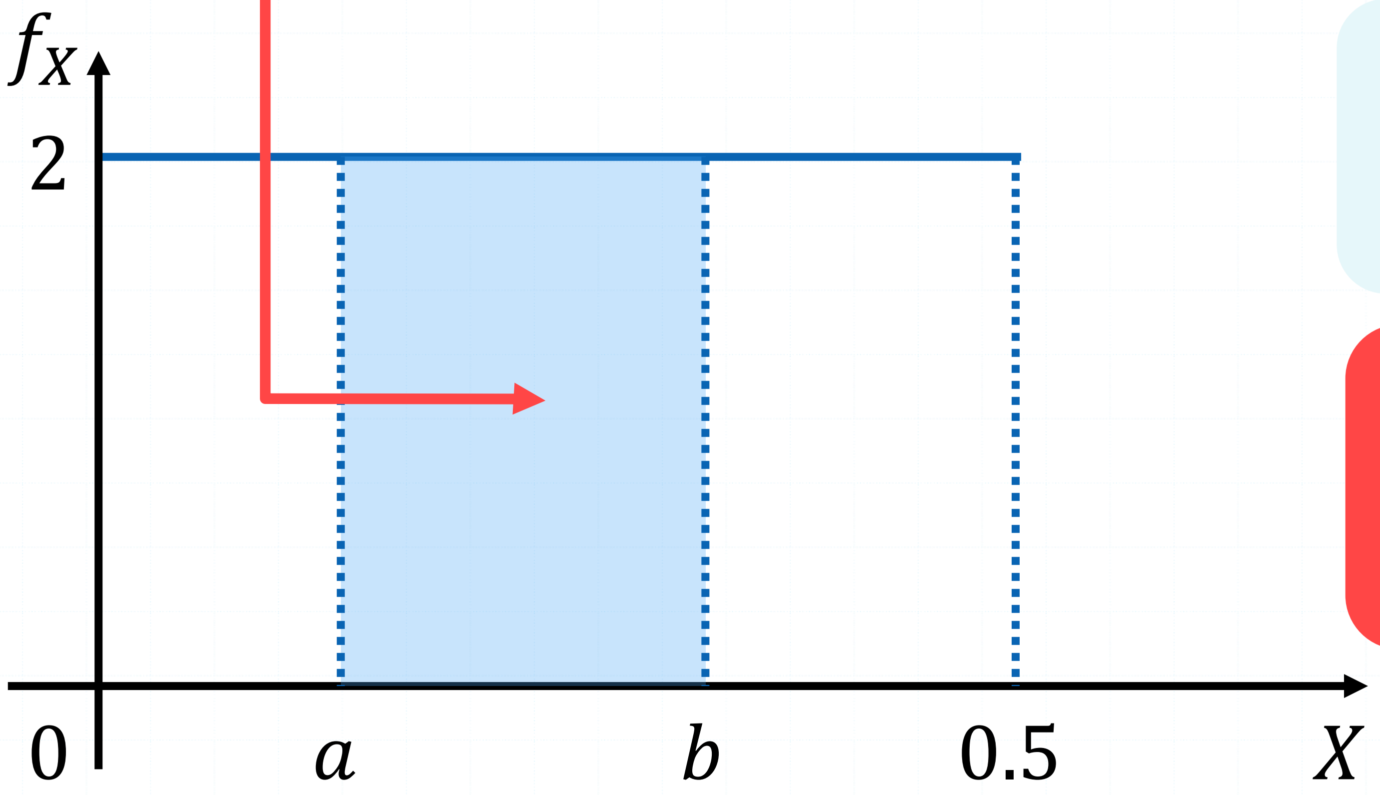
確率密度関数の例

0から0.5の数直線上にランダムに針を落としたときの位置

確率変数 X: $0 \leq X \leq 0.5$

確率密度関数 $f_X(x) = 2\ (0 \leq x \leq 0.5)$ とすると,

$$P(a \leq X \leq b) = \frac{b-a}{0.5} = \frac{\text{区間の幅}}{\text{全体の区間の幅}}$$



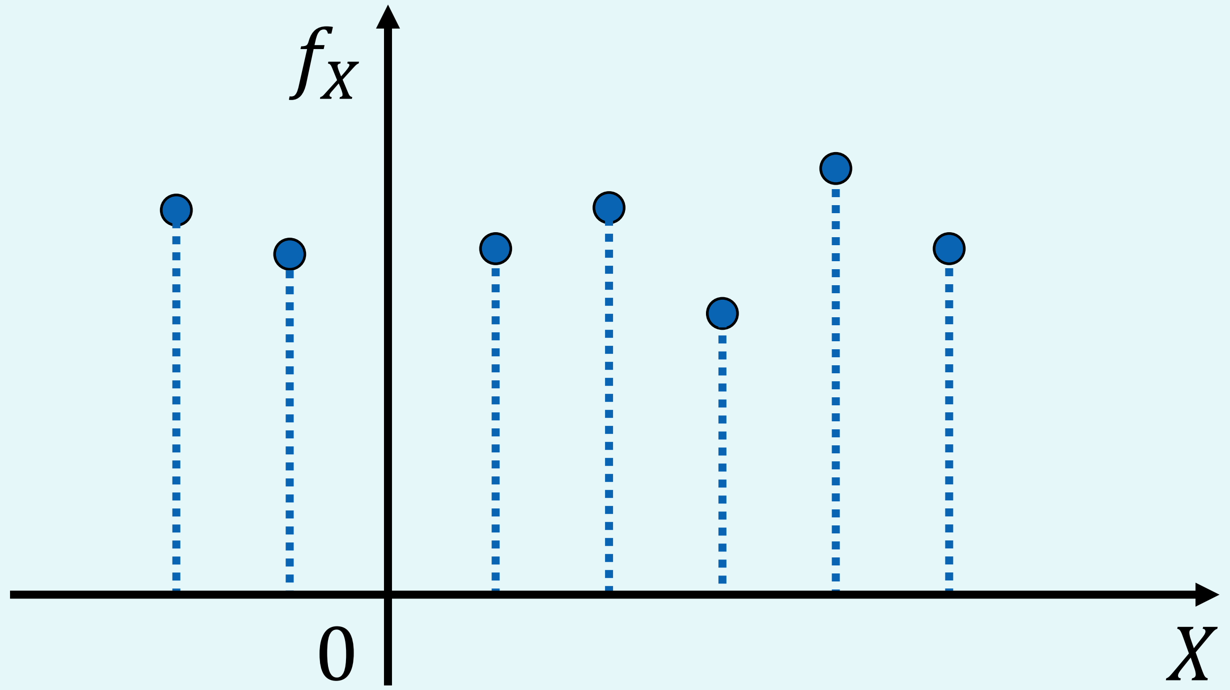
$$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx$$

縦軸の値（f_Xの値）は確率ではない!!

離散型と連続型で
グラフの見方が違うことに
注意しましょう

離散型, 連続型のまとめ

離散型分布

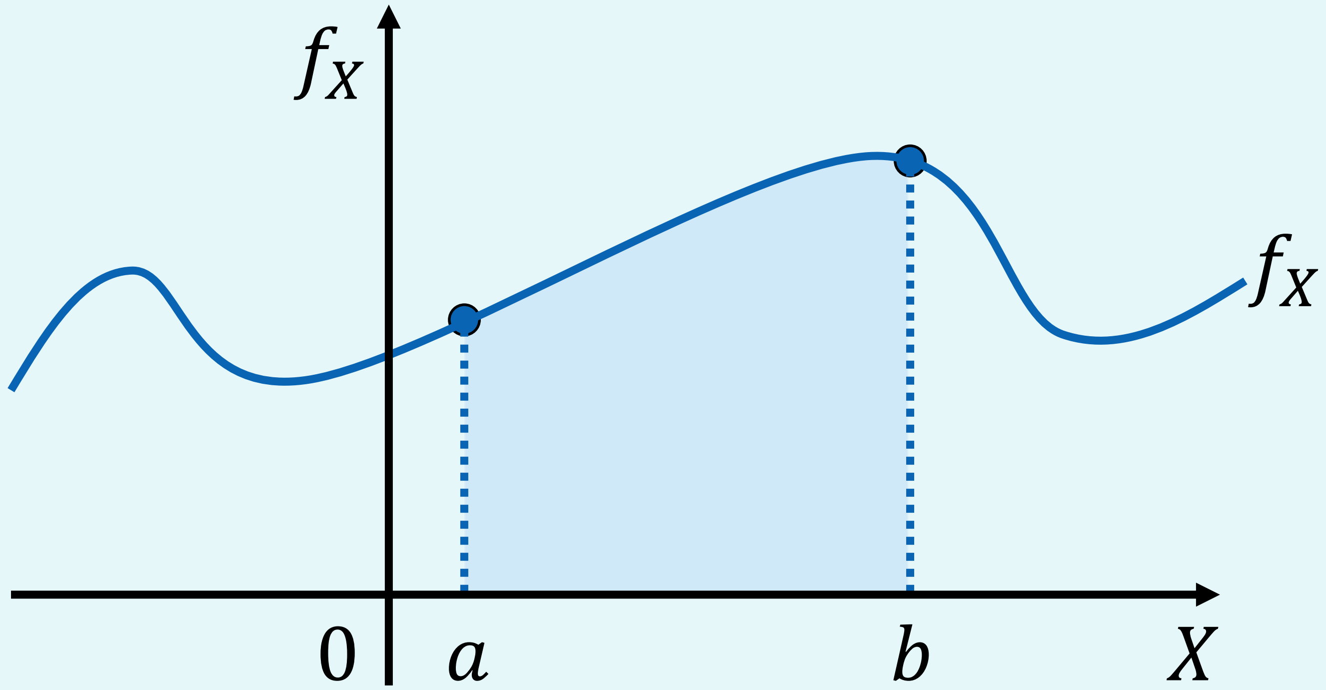


確率関数: $f_X(x_k)$

f_X が**1つの値**の確率

$$f_X(x_k) = P(X = x_k)$$

連続型分布



確率密度関数: $f_X(x)$

f_X **の面積**が**幅**の確率

$$P(a \leq X \leq b) = \int_a^b f_X(x)\,dx$$

離散型確率変数の期待値（平均）, 分散

データの平均値, 分散 (復習)

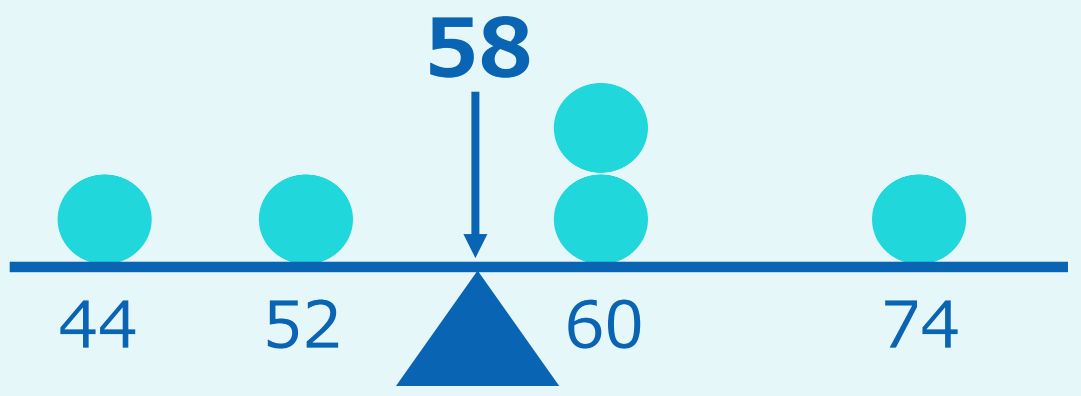
表. 5人の体重データ

人	1	2	3	4	5
体重 (kg)	60	52	44	74	60

平均値

データの重心を表す

$$\frac{60+52+44+74+60}{5} = \mathbf{58}$$



分散

平均値周りのバラつき具合を表す

$$\frac{(60-\mathbf{58})^2+(52-\mathbf{58})^2+(44-\mathbf{58})^2+(74-\mathbf{58})^2+(60-\mathbf{58})^2}{5} = 99.2$$

似たようなものが確率変数でもある!!

離散型確率変数の期待値 (平均)

くじ引き1

	A賞	B賞	C賞
賞金 (円)	1000	500	100
本数 (本)	1	3	6

確率変数 X: 1回くじを引いたときの賞金額 x 円

確率変数 X	1000	500	100	計
確率	1/10	3/10	6/10	1

離散型確率変数 X の期待値 (平均) (Expected value)

$E[X]$ = (確率変数 X の値) × (その確率) $+\cdots$

どれくらいの値を中心にして X の値が起こりやすいかを表す

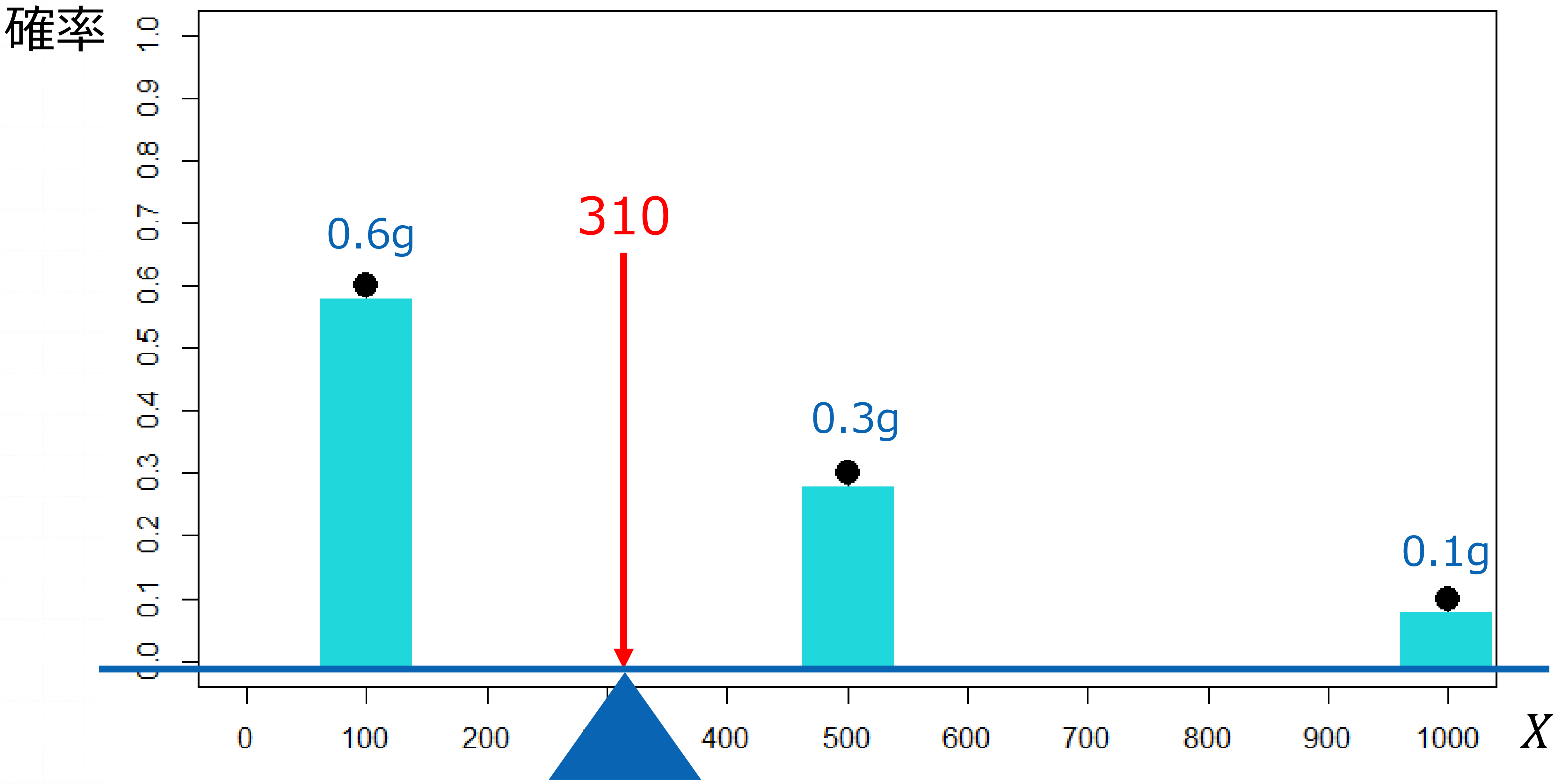
くじ引き1

$$E[X] = 1000 \times \frac{1}{10} + 500 \times \frac{3}{10} + 100 \times \frac{6}{10} = 310$$

何度もくじを引くと
310円周りに賞金が集まる

期待値の視覚的イメージ

確率変数 X	1000	500	100	計
確率	1/10	3/10	6/10	1



! $E[X]$ は確率分布の重心を表す

確率変数の期待値

練習問題 (1) のくじ

	A賞	B賞	C賞
賞金 (円)	1000	500	100
本数 (本)	1	3	6

期待値の計算には,
例えばこんな利用方法が
考えられます

400円で1回だけくじを引けます

お得なら引きたい

期待値 310 ＜ 支払額 400
だから損しそう ⇒ 引かないでおこう

離散型確率変数の分散

離散型確率変数 X の分散

$$
\begin{aligned}
V[X] &= \{(\text{確率変数 } X \text{ の値}) - E[X]\}^2 \times (\text{その確率}) + \cdots \\
&= E[(X - E[X])^2]
\end{aligned}
$$

期待値周りの (X の起こりやすさの) バラつき具合を表す

$V[X]$ が小さいほど期待値に近い X の値が起こりやすい

練習問題 (1) のくじ

$$
\begin{aligned}
V[X] &= (1000 - 310)^2 \times \frac{1}{10} + (500 - 310)^2 \times \frac{3}{10} + (100 - 310)^2 \times \frac{6}{10} \\
&= 84900
\end{aligned}
$$

違いに注意

データの平均値, 分散

今持っている**データそのもの**の特徴を表す

データがないと計算できない

確率変数の期待値 (平均), 分散

確率分布の特徴を表す (確率的な現象・仕組みの特徴)

データは必要ないが, **確率分布**がわからないと計算できない

考え方が似てるから名前も似ているだけ
注目しているものは全く異なる!!

データの平均値, 分散とは
考え方自体は似ていても
異なるものなので注意です!

どのくじを引く?

くじ引き1

	A賞	B賞	C賞
賞金 (円)	1000	500	100
本数 (本)	1	3	6

くじ引き2

	A賞	B賞	C賞
賞金 (円)	1000	500	0
本数 (本)	2	2	6

くじ引き3

	A賞	B賞	C賞
賞金 (円)	1000	400	100
本数 (本)	2	1	7

Q 確率変数の**期待値**と**分散**を計算して判断してみよう!!

どのくじを引く?

くじ引き1

	A賞	B賞	C賞
賞金 (円)	1000	500	100
本数 (本)	1	3	6

確率変数 X: くじ引き1の賞金額 X 円

X	1000	500	100	計
確率	1/10	3/10	6/10	1

A

$E[X] = 310$

$V[X] = 84900$

くじ引き2

	A賞	B賞	C賞
賞金 (円)	1000	500	0
本数 (本)	2	2	6

確率変数 Y: くじ引き2の賞金額 Y 円

Y	1000	500	0	計
確率	2/10	2/10	6/10	1

A

$E[Y] = ??$

$V[Y] = 106000$

くじ引き3

	A賞	B賞	C賞
賞金 (円)	1000	400	100
本数 (本)	2	1	7

確率変数 Z: くじ引き3の賞金額 Z 円

Z	1000	400	100	計
確率	2/10	1/10	7/10	1

A

$E[Z] = ??$

$V[Z] = 126900$

練習問題 (2)

Q 期待値の ?? 部分を求めて，どのくじを引くのがいいか考えよ．

確率変数 X: くじ引き1の賞金額 X 円

X	1000	500	100	計
確率	1/10	3/10	6/10	1

$E[X] = 310$
$V[X] = 84900$

確率変数 Y: くじ引き2の賞金額 Y 円

Y	1000	500	0	計
確率	2/10	2/10	6/10	1

$E[Y] = ??$
$V[Y] = 106000$

確率変数 Z: くじ引き3の賞金額 Z 円

Z	1000	400	100	計
確率	2/10	1/10	7/10	1

$E[Z] = ??$
$V[Z] = 126900$

A

連続型確率変数の期待値, 分散 (参考)

連続型確率変数の期待値, 分散

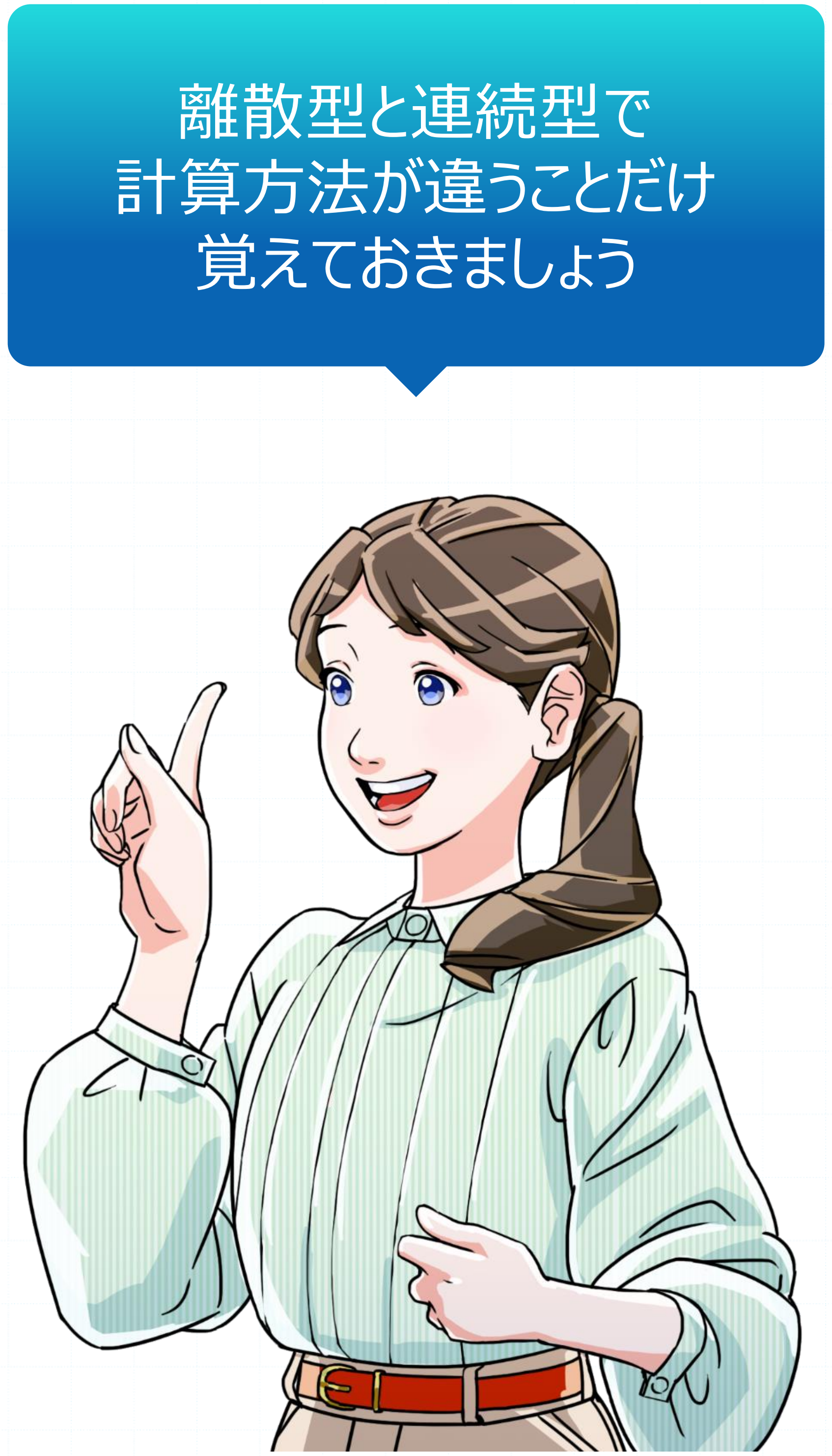
連続型確率変数 X の確率密度関数を $f_X(x)$ とする.

X の期待値 (平均):

$$E[X] = \int x f_X(x)\,dx$$

X の分散:

$$V[X] = E[(X - E[X])^2] = \int (x - E[X])^2 f_X(x)\,dx$$



今日のまとめ

- **確率変数:**
  事象を数字に置き換える変換器
  離散型確率変数 or 連続型確率変数

- **確率分布:**
  確率変数のとる値とその確率の対応関係
  離散型分布, 連続型分布
  確率分布は現象を理解するための道具

- **期待値, 分散:**
  確率分布の特徴を表す指標
  離散型 or 連続型で異なる
  データの平均, 分散とは異なる!!