先週までの「データ解析」

Prologue: データを読み解くとは何なのか
(多変量解析とデータサイエンスと統計学とパターン認識と機械学習とデータマイニング)

DAY-1 6/16 (01)(02) **単回帰: 点群への直線当てはめを”真剣に”考える**
(見えない世界へようこそ)
DAY-2 6/23 (03)(04) **重回帰と線形代数: 回帰の行列計算とその意味**
(データの計算とデータの解釈)

今日の内容

DAY-3 6/30 (05)(06) **重回帰と確率統計: なぜ回帰に確率が必要?**
(推測統計入門:データの向こう側について語るための代償)

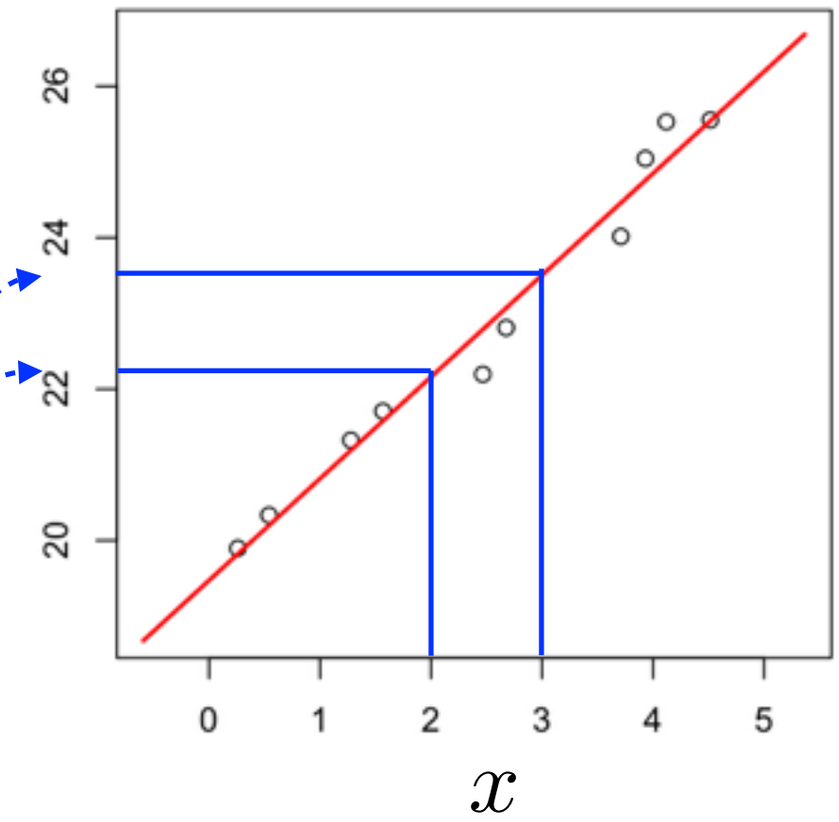
やりたいこと：回帰

$$\hat{y} = 1.343664\, x + 19.474212$$

x	y	$\hat{y}$
0.541	20.336	20.20113442
1.277	21.326	21.19007119
0.26	19.898	19.82356481
3.93	25.043	24.75481204
2.676	22.81	23.06985726
2.466	22.195	22.7876878
1.566	21.708	21.57839012
4.119	25.528	25.00876455
4.515	25.554	25.54085554
3.709	24.015	24.45786227
2.0	NA	22.16154
3.0	NA	23.5052

手元にあるデータ

手元にないデータ



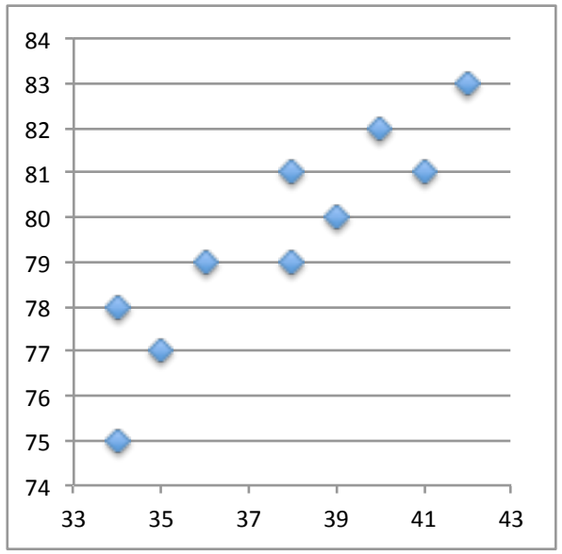
データ点にモデル式を当てはめることで
データにない点について何か語る

$$\boldsymbol{X} = \begin{bmatrix} 0.541 & 1 \\ 1.277 & 1 \\ 0.26 & 1 \\ 3.93 & 1 \\ 2.676 & 1 \\ 2.466 & 1 \\ 1.566 & 1 \\ 4.119 & 1 \\ 4.515 & 1 \\ 3.709 & 1 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 20.336 \\ 21.326 \\ 19.898 \\ 25.043 \\ 22.81 \\ 22.195 \\ 21.708 \\ 25.528 \\ 25.554 \\ 24.015 \end{bmatrix}$$

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \begin{bmatrix} 1.343664 \\ 19.474212 \end{bmatrix}$$

「回帰分析」に至る「シチュエーション」と「気持ち」

- ワイシャツメーカーが「首回り」と「腕の長さ」を**10人**で調べたら「関係」がありそうだった。この「関係」を予測に役立てたい。
- つまり本当の関心は**この特定の10人の傾向ではなく**「首回り」と「腕の長さ」に関する**一般的な傾向(法則性)**



	首回り	腕の長さ
被験者1	38	81
被験者2	40	82
被験者3	34	78
被験者4	41	81
被験者5	34	75
被験者6	38	79
被験者7	42	83
被験者8	36	79
被験者9	35	77
被験者10	39	80

- **疑問：**
- **別の10人で調べたらこの傾向はどれくらい変わり得る？？**
- **この10人に関する調査結果から、一般に法則性があることをどれくらいの確度で言うことができる？？**

そんなこと聞かれても10人のデータだけしかないんだから、
何も言えない…😫

→ **確率の導入**

つまり、手元に得られているデータの背後にある**未知なる法則性**を「仮定」し、それをデータから当てる問題として考えなおす。

今日のはなし (まずは単変量で！来週、多変量版＋α)

混乱したらここ(初心)へ戻り、**太字**の概念を理解したか確認！

[午前]
- **確率**ってそもそも何なの？
- 3つの道具：**確率変数**、**確率分布**、**期待値**
- 確率から統計へ：**母集団**と**標本**, **統計量**と**標本分布**
- **正規分布**とその性質

[午後]
- 正規分布の兄弟 (**カイ二乗分布**, **t分布**)
- 確率を導入しないとできないこと：**区間推定**と**仮説検定**
- **正規線形モデル**と**回帰係数の検定**
- 回帰係数・母回帰の区間推定：**予測区間**と**信頼区間**

復習：単回帰を考えて得られたこと①

- 単変量の場合でも**行列やベクトル**を使えば問題が美しく解けるのを見ることができる！（線形代数との接点）
- 単にもっとも当てはまりの良い回帰の係数(傾きと切片)を計算すれば良いだけなら「線形代数」があればよくて「確率統計」の道具は要らない。

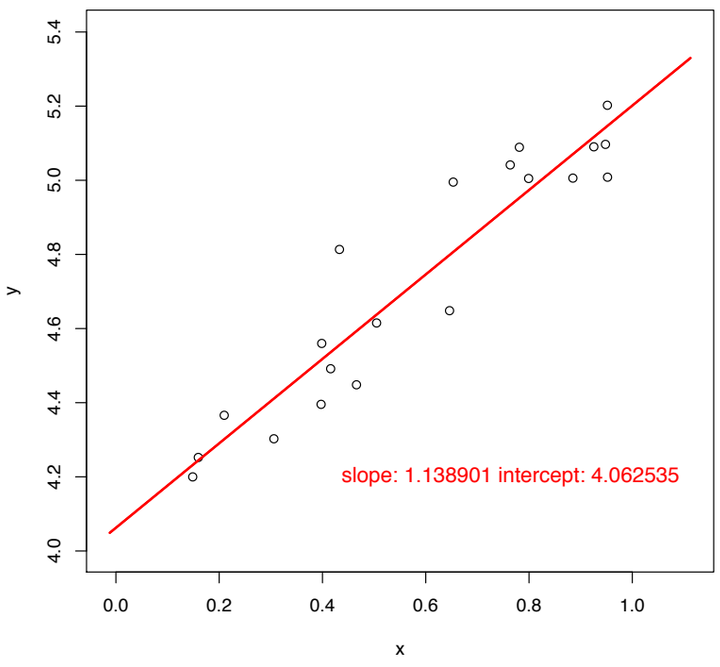
$\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ や $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$ の意味は？

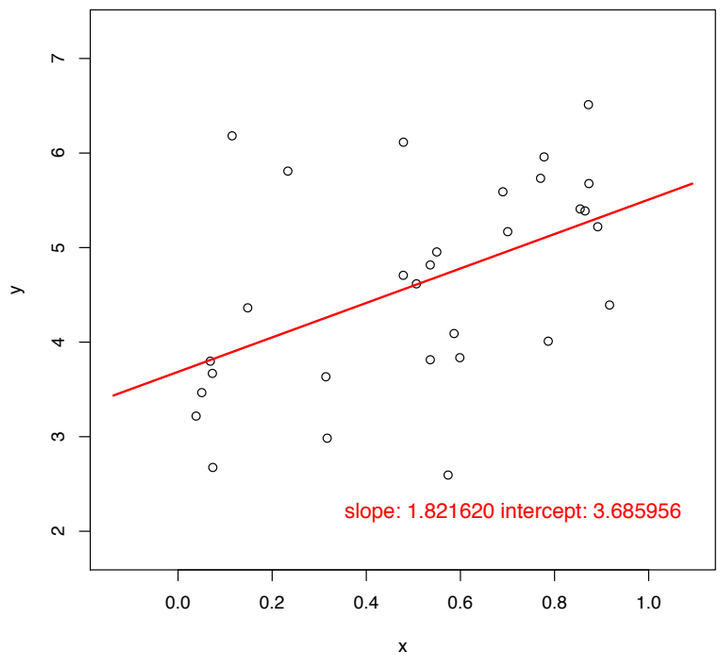
→ 回帰計算の意味を先週の後半でもう少し掘り下げました。
ベクトル・行列表記が活きる多変量解析計算の真骨頂！

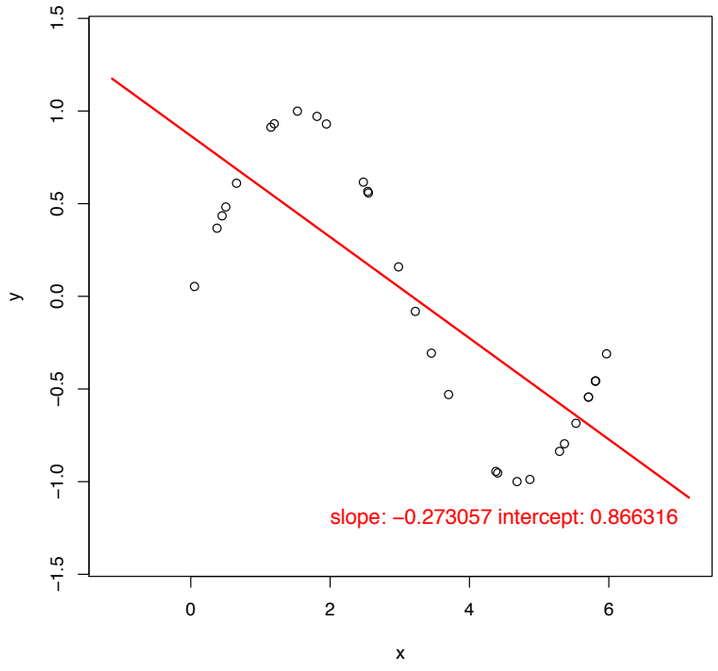
復習：単回帰を考えて得られたこと②

$(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ が存在するならどんなに変なデータでも

線形回帰の「計算」は常に可能であることに注意





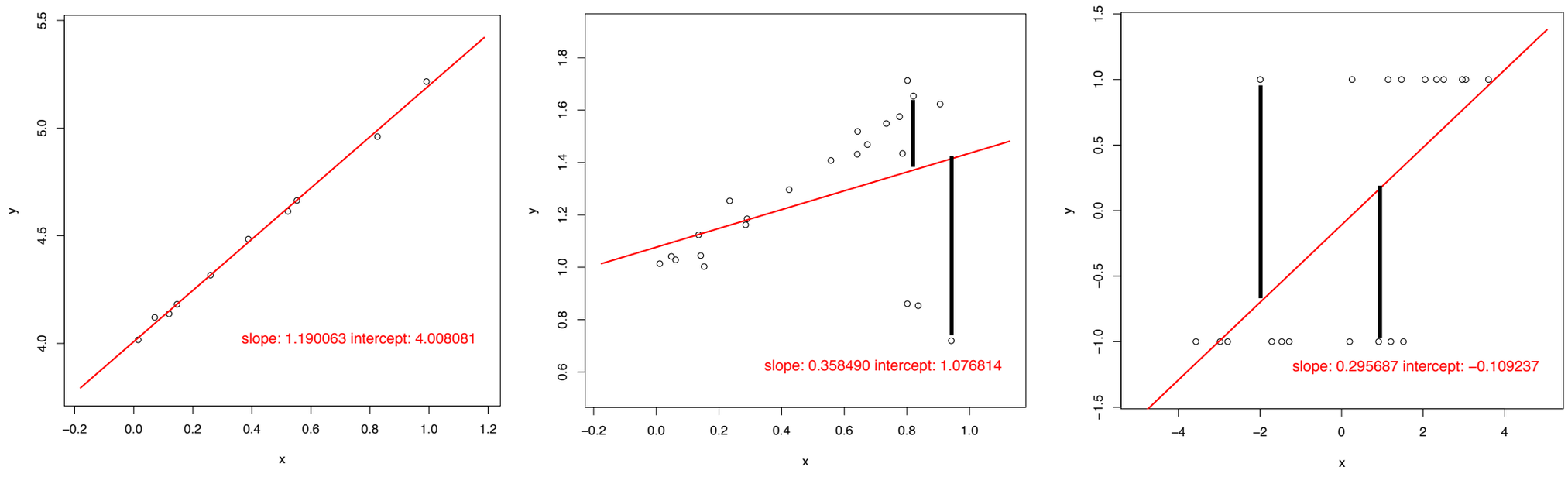


じゃあ良し悪しはどうやってわかるの？

→ ここで確率統計が登場！今週のメインテーマ

復習：単回帰を考えて得られたこと③

二乗誤差の「総和」を小さくするので、手元に既にある
データであっても予測値がかなり悪くなり得る。

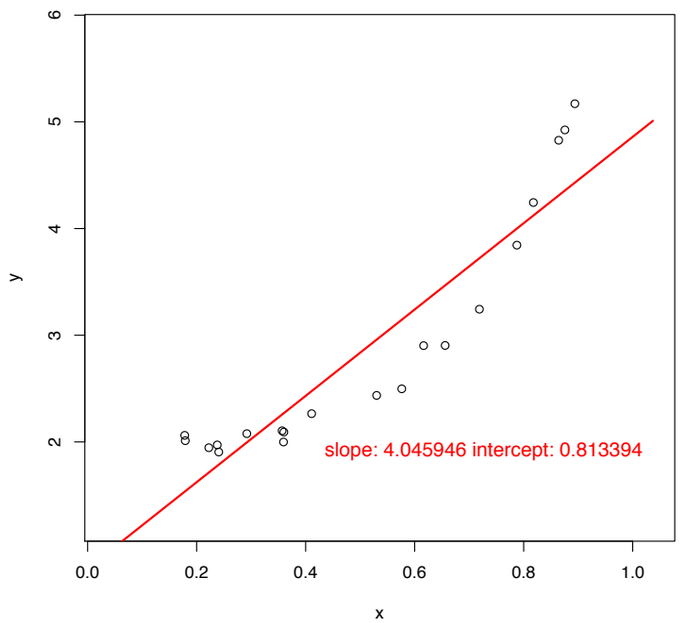


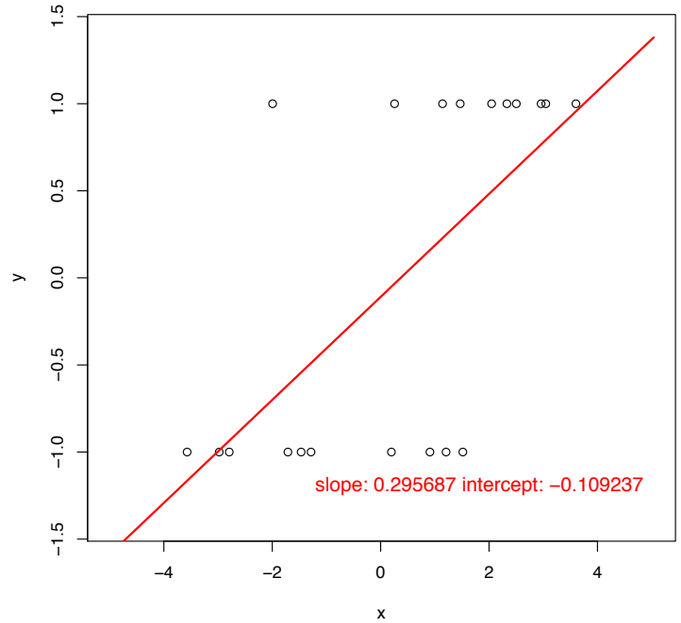
$\varepsilon_i = \hat{y}_i - y_i$ に対して、なぜ $|\varepsilon_1| + |\varepsilon_2| + \cdots + |\varepsilon_n|$
ではなく、二乗した $\varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_n^2$ を最小に？

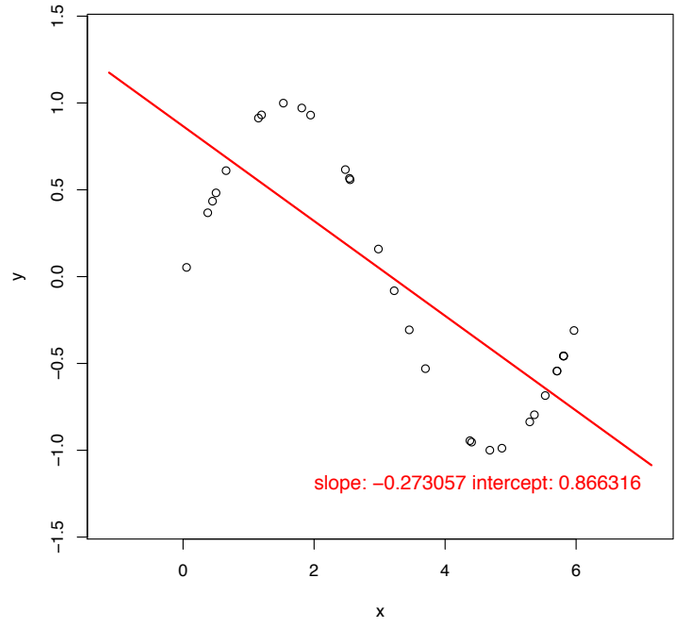
→ これにも確率統計が必要。今週のテーマの一つ

復習：単回帰を考えて得られたこと④

- 手法の良し悪しは**モデル式がデータに適合しているかどうか**で決まる！データに依らず、ある方法が良いとか悪いとか言うことはできない。







直線ではなくて曲線を当てはめるにはどうする？

→ 先週取り上げました。一般には機械学習の領域へ。

準備事項

統計やるのに必須な３つの重要概念をつかむ。

１．確率変数と確率分布

２．(確率変数の)期待値

３．i.i.d.

★ 確率変数

ある確率に従って定義域内の様々な値をとるような特殊な変数

確率変数 X と書いたら、時には $X = 3.5$ だったり

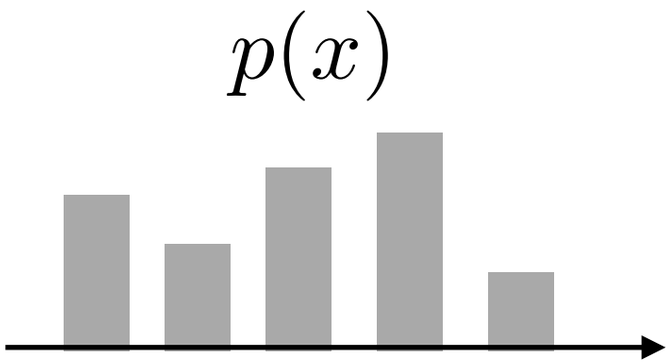
$X = -0.2$ だったりする。

どの値をどれくらいの率で取るかは
その**確率分布**で定まっている

★ 確率分布

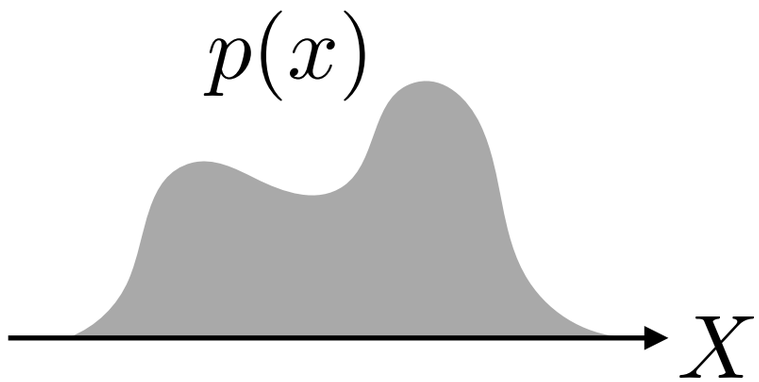
確率変数 X がどの値を取るかの起こりやすさを記述

X が離散の場合



全可能性を
網羅すると1

X が連続の場合



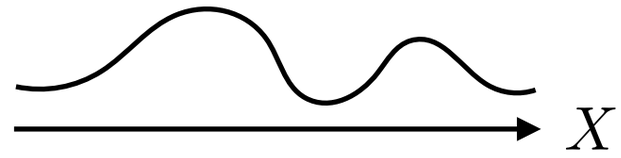
★ 確率変数とは!?

確率変数 X

＝ **ある確率に従って定義域内の様々な値をとる**ような特殊な変数

確率変数 X と書いたら、時には $X = 3.5$ だったり

$X = -0.2$ だったりする。



どの値をどれくらいの率で取るかは
確率で定まっている

数学的には X と書いたら**常に同じものを差してほしい**ので
厳密にこういうのを定めるのは結構手間がかかる…

数学的には普通、X と書いて、(ランダムに)どれかが起こる対象を全て考えた集まり Ω の各要素に実数を割り当てる**関数**と見なす
(関数なら引数によって値が変わって良いので)

標本空間　$\Omega = \{$  $\}$

$X : \Omega \to \mathbb{R}$　　$X($  $) = 1$　$X($  $) = 2$　$X($  $) = 3$

(関数として)　$X($$) = 4$　$X($$) = 5$　$X($$) = 6$

これで「毎度、値が変わる」点はOK。

でもこれだと「出来事」を引数とする変な実関数を定義しただけ。
真のポイントはこの標本空間上で「確率」を測る操作の定義。

★ 確率とは!?

Ω の部分集合の集まり $\mathcal{F}$ の要素に0～1の実数を割当

標本空間　$\Omega = \{$  $\}$

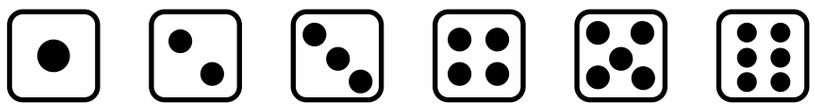
“奇数が出る”　$\{$  $\} \in \mathcal{F}$

確率　$P(\{$  $\}) \leftarrow$ [0,1]に値をとる数字

※標本空間が連続の場合、部分集合の全ての集まりだと変な確率を定義できる対象も入ってしまうので、「確率」が満たしてほしい性質をいつも P が満たすように $\mathcal{F}$ の定義を少しいじる必要がある。

($\mathcal{F}$ をσ-fieldというやつにする)

★ 確率変数を使って確率を考えるメリット

標本空間　$\Omega = \{$  $\}$

確率変数　$X : \Omega \to \mathbb{R}$

出来事 $\omega \in \Omega$ に依って値 $X(\omega) \in \mathbb{R}$ が決まる

確率　$P(\{$  $\}) = P(X \in \{1, 3, 5\})$

- 「出来事(左)」は表記に難ありだが、「実数(右)」は表記が楽
- 「値は試行ごとに変わるけど、ともかく何かの実数値」という形として、ランダムな過程を含む対象をうまく定義した！

以後、確率変数 X が出て来たら、その背後に標本空間 Ω と確率 P が何か既に定まっている点を思い出そう！

★ 厳密にやるための蛇足

厳密には、確率を定めるには前述の $\mathcal{F}$ も定めないといけなくて
これを使って確率の測り方(確率測度) $P : \mathcal{F} \to [0, 1]$ が定まる。
また関数としての確率変数 X もこの $\mathcal{F}$ と整合する必要がある。
$\Rightarrow$ 確率空間 $(\Omega, \mathcal{F}, P)$ 上の $\mathcal{F}$-可測関数を確率変数と言う。

★ 「確率変数」と「確率変数が取る値」の区別に注意

3回の試行で確率変数 X が取った値が各々 x_1, x_2, x_3 であった。
とか言うとき、x_1, x_2, x_3 は単なる実数値で一方 X は確率変数
(試行毎に値が変わる変数)だ、という差に注意。

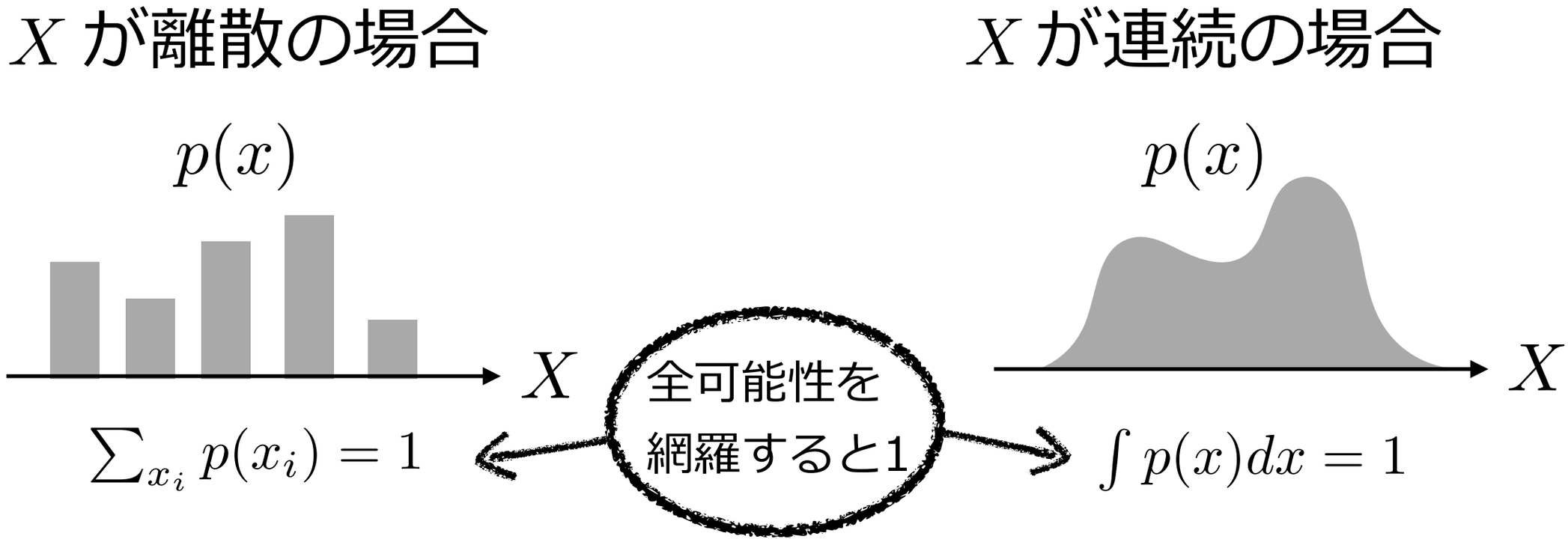
例えば、$x_1 X - x_2$ は新しい確率変数になるが、確率的に変動
するのは X のところだけ！

★ 確率分布

確率変数 X がどの値を取るかの起こりやすさを記述

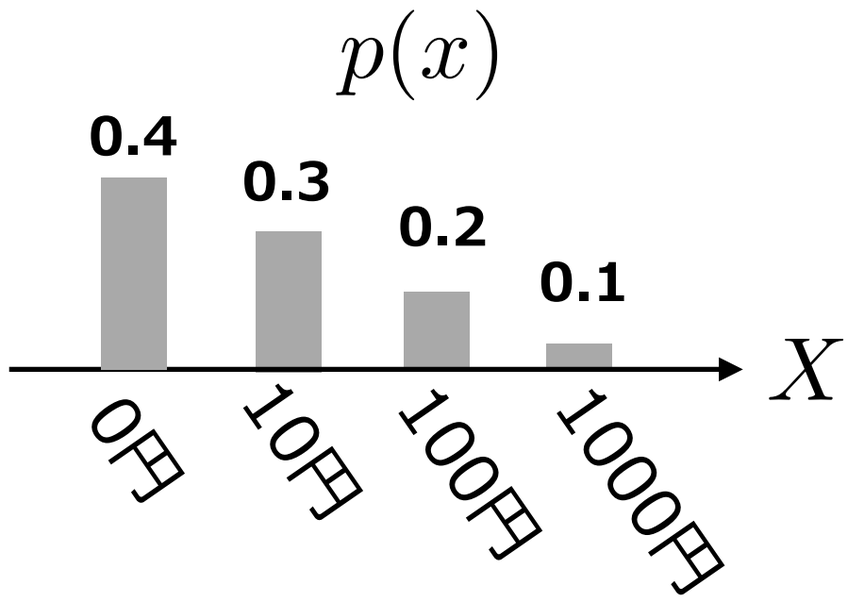
→ 確率測度 P から定まる実数空間 $\mathbb{R}$ 上の確率測度

…なのだが、とりあえずは「密度関数」をイメージすればよい



★ 期待値(expected value/expectation)とは!?

何度も何度も試行したとき確率変数 X が平均的に取る値

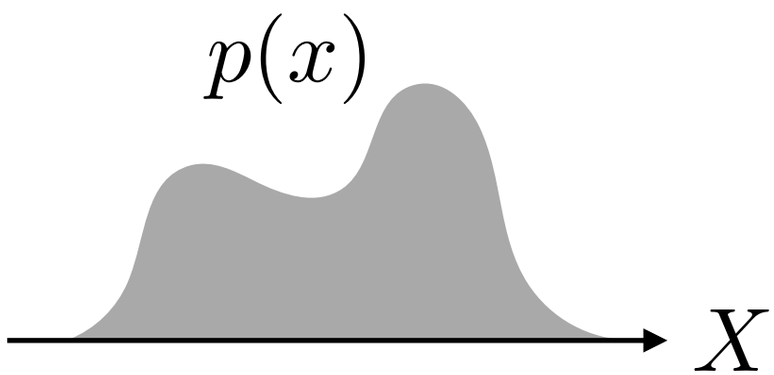


150円のたからくじ1つの当たり額X

期待値 (平均いくら当たるか) = 123円

$$0.4 \times 0 + 0.3 \times 10 + 0.2 \times 100 + 0.1 \times 1000$$

$$= \sum_{x_i} p(x_i) x_i$$ ← **確率による重みつき平均値**



連続の場合、積分になるけど意味は同じ

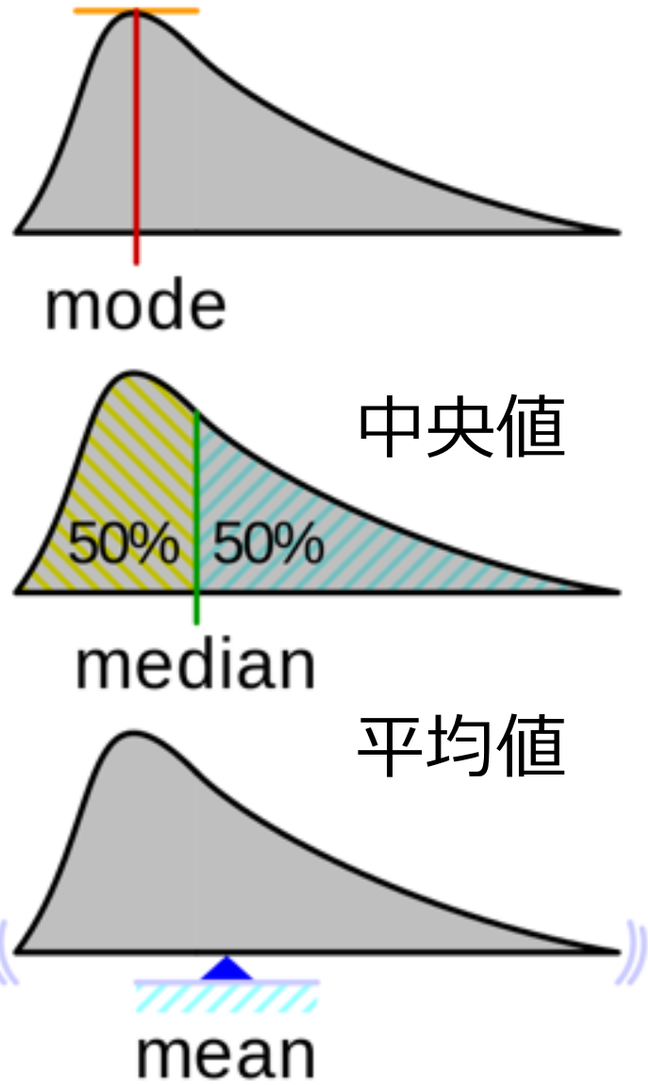
$$\int_x p(x) x dx$$

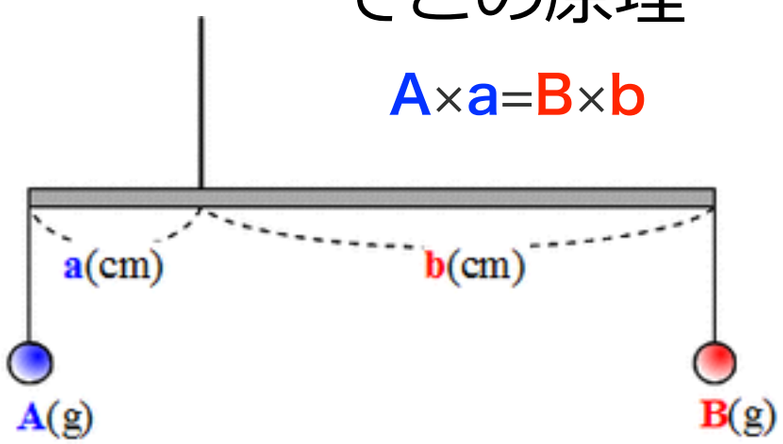
★ 期待値の計算

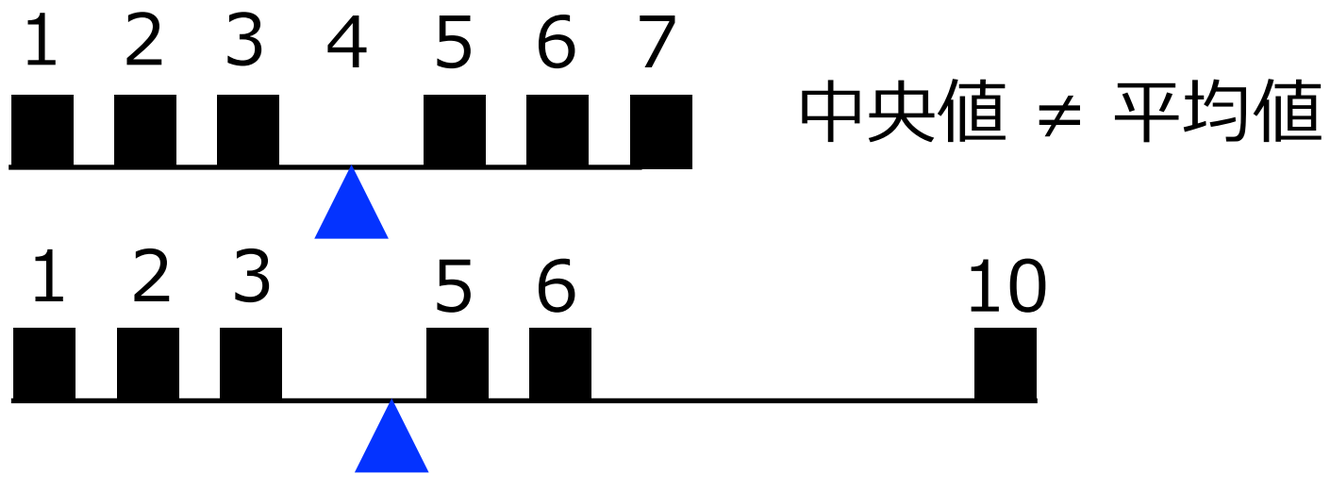
確率分布を考慮した「平均(mean)」

$\mathbb{E}\{\cdot\}$ と書く。

$$\mathbb{E}\{X\} := \int_x x \cdot p(x) dx$$

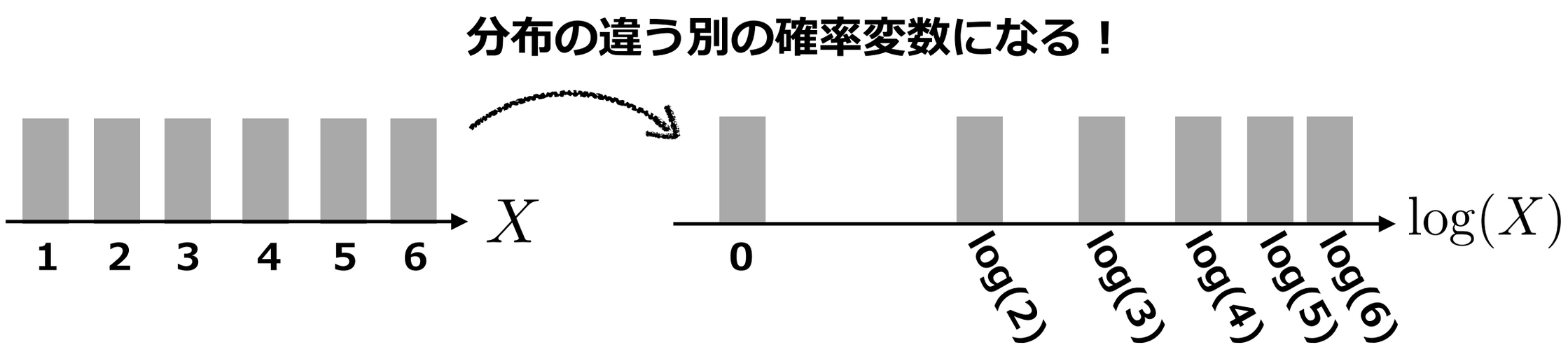






★ 期待値の計算

$$\mathbb{E}\{f(X)\} = \int_x p(x) f(x) dx$$



$$\mathbb{E}\{X\} = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\mathbb{E}\{\log(X)\} = \frac{0+\log(2)+\log(3)+\log(4)+\log(5)+\log(6)}{6} = \frac{\log(720)}{6}$$

$$\approx 1.096542$$

★ 「確率変数」と「確率変数が取る値」の区別に注意

3回の試行で確率変数 X が取った値が各々 x_1, x_2, x_3 であった。
とか言うとき、x_1, x_2, x_3 は単なる実数値で一方 Xは確率変数
(試行毎に値が変わる変数)だ、という差に注意。

例えば、$x_1X - x_2$ は新しい確率変数になるが、確率的に変動するのは Xのところだけ！

$$\mathbb{E}\{x_1X - x_2\} = \int_z (x_1z - x_2) \cdot p(z)dz$$

$$= x_1 \cdot \left(\int_z zp(z)dz \right) - x_2$$

$$= x_1 \cdot \mathbb{E}\{X\} - x_2$$

期待値は確率変動するところへ作用

★ 期待値の超基本的な性質 その１：線形性

$$\mathbb{E}\{aX + bY + c\} = a\mathbb{E}\{X\} + b\mathbb{E}\{Y\} + c$$

証明：期待値計算は積分なので。

$$\mathbb{E}\{aX + bY\} = \int_{\omega\in\Omega} (aX(\omega) + bY(\omega) + c)P(d\omega)$$

$$= a\int_{\omega\in\Omega} X(\omega)P(d\omega) + b\int_{\omega\in\Omega} Y(\omega)P(d\omega) + c\int_{\omega\in\Omega} P(d\omega)$$

例)さきほどのサイコロの出る目 X に対して

$$\mathbb{E}\{3\log(X) - 1.5X\} = 3\mathbb{E}\{\log(X)\} - 1.5\mathbb{E}\{X\}$$

$$= 3\cdot\frac{\log(720)}{6} - 1.5\cdot 3.5 \approx -1.960374$$

★ 期待値の超基本的な性質 その２：独立なとき

確率変数 X と Y は次を満たすとき独立という

$$P(X \in A \text{かつ} Y \in B) = P(X \in A)P(Y \in B)$$

確率密度で言うと、$p(x,y) = p(x)p(y)$

> X と Y が独立のとき、$\mathbb{E}\{XY\} = \mathbb{E}\{X\}\mathbb{E}\{Y\}$

例)さきほどのサイコロを2回ふって各々 X, Y が出たとして

$$\mathbb{E}\{X\log(X)\} = \mathbb{E}\{\log(X^X)\} = \frac{\log(2^2 \cdot 3^3 \cdot 4^4 \cdot 5^5 \cdot 6^6)}{6}$$

$$\neq \mathbb{E}\{X\}\mathbb{E}\{\log(X)\} = 3.5 \cdot \frac{\log(720)}{6}$$

$$\mathbb{E}\{X\log(Y)\} = \mathbb{E}\{X\} \cdot \mathbb{E}\{\log(Y)\} = 3.5 \cdot \frac{\log(720)}{6} \approx 3.837897$$

★ 注意：期待値は確率変数ではなく単なる値です！

$\mathbb{E}\{X\}$ は**単なる値**で確率的に変動しないので
(つまり、確率変数ではないので)その値のまま

例　$\mathrm{var}(X) := \mathbb{E}\{(X - \mathbb{E}(X))^2\}$

$$= \mathbb{E}\{X^2 - 2\boxed{\mathbb{E}(X)} \cdot X + (\boxed{\mathbb{E}(X)})^2\}$$

$$= \mathbb{E}\{X^2\} - (\mathbb{E}(X))^2$$

紛らわしい場合、$\mathbb{E}\{X\} = \mu$　と置いてしまって

$$\mathbb{E}\{(X-\mu)^2)\} = \mathbb{E}\{X^2 - 2\mu X + \mu^2)\}$$

$$= \mathbb{E}\{X^2\} - 2\mu\mathbb{E}\{X\} + \mu^2$$

$$= \mathbb{E}\{X^2\} - 2\mu^2 + \mu^2 = \mathbb{E}\{X^2\} - \mu^2$$

★ サイコロの例をRで検証

```
> x <- floor(runif(100000,min=0,max=6))+1
> y <- floor(runif(100000,min=0,max=6))+1
> table(x)
x
    1     2     3     4     5     6
16880 16605 16717 16484 16656 16658
> table(y)
y
    1     2     3     4     5     6
16463 16740 16750 16672 16636 16739
```

xとyは独立にサンプリング

←100000個の乱数
(1～6を等確率で取る)

←各々の値の頻度

```
> mean(x)
[1] 3.49405
> mean(y)
[1] 3.50495
> mean(log(x))
[1] 1.093808
> mean(log(y))
[1] 1.098842
> mean(3*log(x)-1.5*x)
[1] -1.959651
> mean(3*log(y)-1.5*y)
[1] -1.960899
```

期待値を
標本平均値
で推定

```
> mean(x*log(x))
[1] 4.826394
> mean(x)*mean(log(x))
[1] 3.82182

> mean(x*log(y))
[1] 3.84523
> mean(x)*mean(log(y))
[1] 3.839409
> mean(y*log(x))
[1] 3.837266
> mean(y)*mean(log(x))
[1] 3.833743
```

独立でない
ので異なる

独立なので
同じ

★ i.i.d. (独立に同一分布に従う)

確率変数 $X_1, X_2, \ldots, X_n$ が各々独立であり、

全く同じ分布 P に従うとき、i.i.d.であると言う。

independent and identically distributed (i.i.d.)

例　　$X_1, X_2, X_3, X_4, X_5 \overset{\text{iid}}{\sim} P$

$$\mathbb{E}\{X_i\} = \mu, \quad \mathbb{E}\{(X_i - \mu)^2\} = \sigma^2$$

とすると

$$\mathbb{E}\{X_1 + X_2 + X_3 + X_4 + X_5\} = 5 \cdot \mu$$

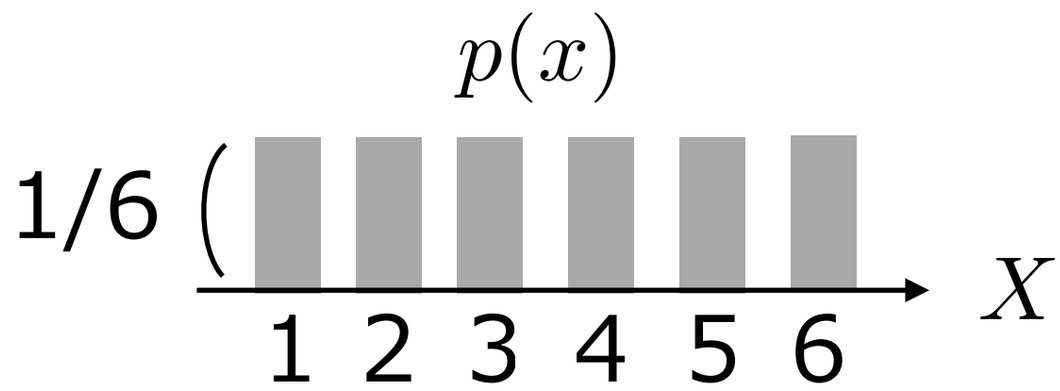
$$\mathbb{E}\left\{\frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}\right\} = \mu$$

```
> x <- floor(runif(100000,min=0,max=6))+1
> y <- floor(runif(100000,min=0,max=6))+1
> table(x)
x
    1     2     3     4     5     6
16880 16605 16717 16484 16656 16658
> table(y)
y
    1     2     3     4     5     6
16463 16740 16750 16672 16636 16739
```

←100000個の乱数
(1～6を等確率で取る)

←各々の値の頻度

xの要素値およびyの要素値は次の確率分布 p(x) に従う100000個のi.i.d.確率変数の実現値と見なせる



推測統計入門

推定量の期待値という考え方にまずは慣れる。

１．母集団と標本統計量

２．「標本平均」という推定量の期待値と分散

３．「標本分散」という推定量の期待値と分散

★ 母集団と標本統計量

観測される”例”の世界での
Xの実現値を**標本(サンプル)**と言う

観測されない背後の確率規則
の世界での可能なXの全体を
母集団と言う

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} P$$

【観測できる量で作った量=標本統計量】

標本平均

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

標本分散

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2$$

関心：
どれくらい当てられるのだろう？

【観測できない量】

$X \overset{\text{iid}}{\sim} P$　に対して

母平均

$$\mu = \mathbb{E}\{X\}$$

母分散

$$\sigma^2 = \mathbb{E}\{(X-\mu)^2\}$$

★ 標本統計量を推定量として考える

標本統計量を使って、母集団に関する未知量を推定するとき、**「推定量(estimator)」**と言う。

手元に得られる量

標本平均　　“母平均の推定量”

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

標本分散　　“母分散の推定量”

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2$$

当たるかな?

当たるかな?

手元に得られない未知量

母平均

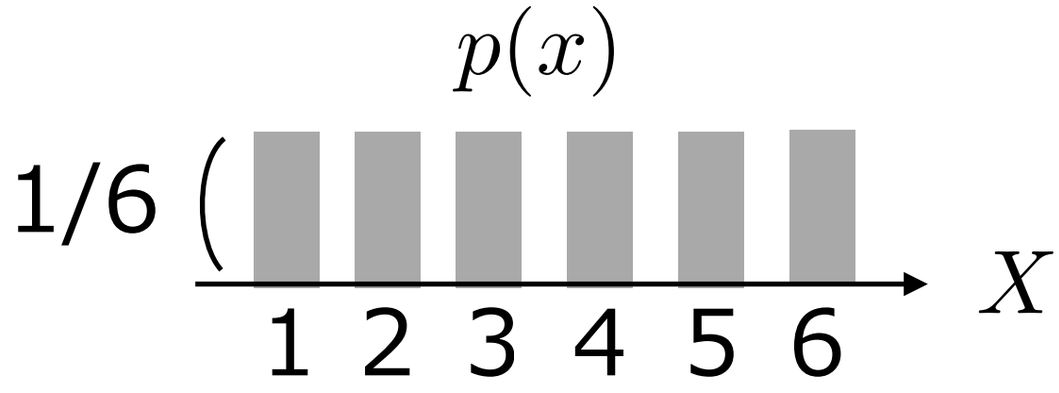
$$\mu = \mathbb{E}\{X\}$$

母分散

$$\sigma^2 = \mathbb{E}\{(X-\mu)^2\}$$

例)

いま右の分布 p(x) が
あるが未知だと仮定



ただしこの分布から次の標本が20個得られている。
(i.i.d.とする)

１５３４５３４２１２５１５３２１５２４４

この20個の標本の平均値 3.1 (標本平均)

↕

真の平均値 (期待値) は 3.5 (母平均)

★ 推定量は確率変数である。

推定量の例)

標本平均

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

標本分散

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2$$

これらは今あるサンプル(標本)の値 $X_1, X_2, \ldots, X_n$ に依存している。

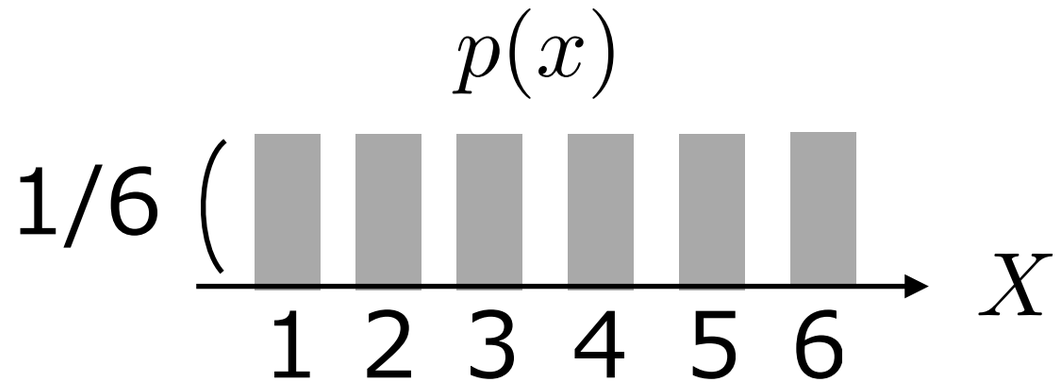
⇨ たまたま得られたn個の値に依存する**確率変数**！

標本分布 = 標本統計量の確率分布

「n個の標本を得る」ことを何度も何度も繰り返すと値が変わる

⇨ 確率変数としての**確率分布**を持つ：期待値や分散なども計算できる

いま右の分布 p(x) が
あるが未知だと仮定



ある20個の標本(i.i.d.)　標本平均 3.1

1 5 3 4 5 3 4 2 1 2 5 1 5 3 2 1 5 2 4 4

別の20個の標本(i.i.d.)　標本平均 3.15

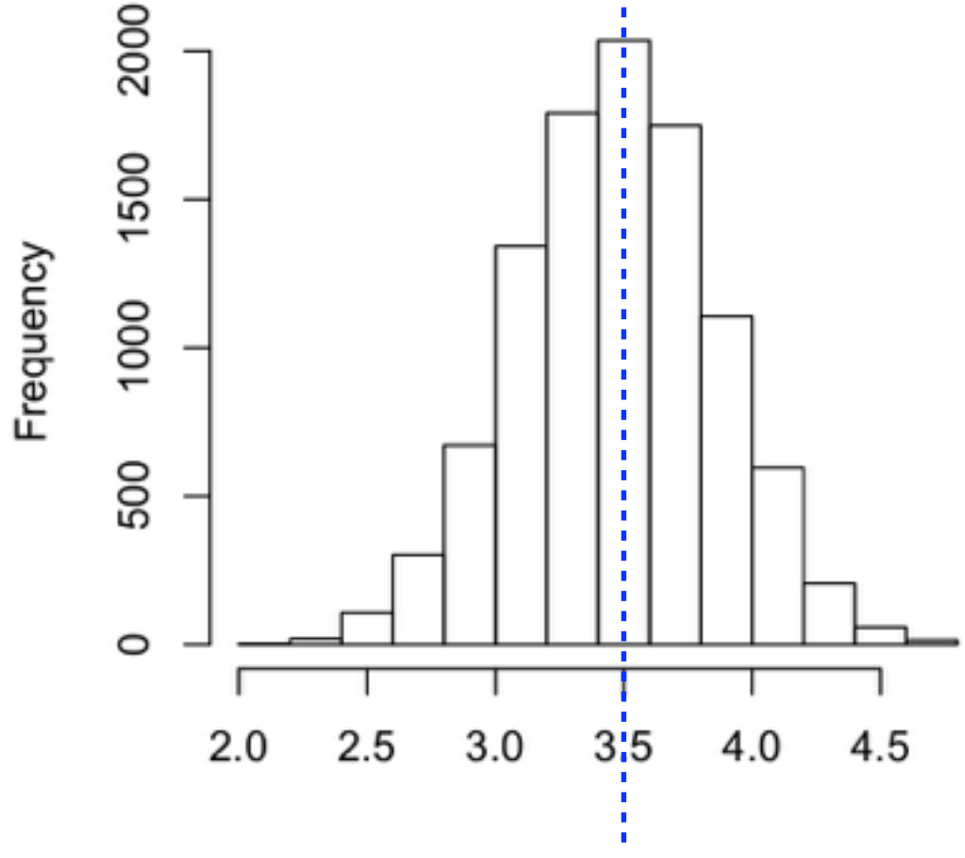
2 2 4 4 2 2 1 3 6 1 1 4 5 1 3 3 5 3 5 6

別の20個の標本(i.i.d.)　標本平均 3.65

4 2 6 3 3 5 5 4 5 2 6 1 2 3 5 5 1 3 6 2

⋮

10000回やってみた場合の
標本平均のヒストグラム



母平均 3.5

★ 母平均の推定量としての標本平均

標本平均の期待値

$$\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}X_i\right\}=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\{X_i\right\}=\mu$$

期待値が求めたいものと一致する推定量を
不偏推定量と言う。(何度も試行をすれば
平均的には当たるという意味で好ましい性質)

標本平均の分散

$$\mathbb{E}\left\{\left(\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)-\mu\right)^2\right\}=\mathbb{E}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\right)^2\right\}$$

i.i.d.変数の独立性より
$i\neq j$ **ならば**
$\mathbb{E}\{(X_i-\mu)(X_j-\mu)\}$
$=\mathbb{E}\{(X_i-\mu)\}\mathbb{E}\{(X_j-\mu)\}=0$
とクロスタームは消える

$$=\mathbb{E}\left\{\frac{1}{n^2}\left(\sum_{i=1}^{n}(X_i-\mu)\right)\left(\sum_{j=1}^{n}(X_j-\mu)\right)\right\}=\frac{1}{n^2}\mathbb{E}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i-\mu)(X_j-\mu)\right\}$$

$$=\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left\{(X_i-\mu)^2\right\}=\frac{n\sigma^2}{n^2}=\frac{\sigma^2}{n}$$

n→無限大で分散→0なので、nが大きい程、求めたいものに近い値が得られる

★ 標本平均の標本分布

(今日は証明しないが)

もしi.i.d.な $X_1, X_2, \cdots, X_n$ が平均μ,分散σ_2の**正規分布**に従うとき、その線形和 $a_1X_1+a_2X_2+\cdots+a_nX_n$ は**正規分布**に従う。

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2) \Rightarrow \bar{X} = \frac{X_1 + \cdots + X_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

★ 中心極限定理

さきほど一様分布のとき見たように実は標本平均の分布は、X_iの分布によらず、(近似的に)**正規分布**に従う。

★ 母分散の推定量としての標本分散

これを考えるために $\mathbb{E}\{X\} = \mu$ および

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \Leftrightarrow \quad n \cdot \bar{X} = \sum_{i=1}^{n} X_i$$

と置き、以下の量を考える。

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X})^2 \\
&= \sum_{i=1}^{n}\{(X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2\} \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + n \cdot (\bar{X} - \mu)^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - n \cdot (\bar{X} - \mu)^2
\end{aligned}$$

ここで、先の標本平均の分散の計算より

$$\mathbb{E}\{(\bar{X}-\mu)^2\} = \frac{\sigma^2}{n}$$

従って、

$$\mathbb{E}\left\{\sum_{i=1}^{n}(X_i-\bar{X})^2\right\} = \mathbb{E}\left\{\sum_{i=1}^{n}(X_i-\mu)^2\right\} - n\cdot\mathbb{E}\left\{(\bar{X}-\mu)^2\right\}$$

$$= n\cdot\sigma^2 - n\cdot\frac{\sigma^2}{n} = (n-1)\sigma^2$$

$$\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right\} = \frac{n-1}{n}\sigma^2$$

←**単純な標本分散は母分散の不偏推定量にならない！**

(ただしnが大なら大体OK?)

$$\mathbb{E}\left\{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right\} = \sigma^2$$

←n-1で割ると不偏

★ 母分散の不偏推定量と自由度

$$\mathbb{E}\left\{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right\}=\sigma^2$$

母分散の不偏推定量を作るには**nじゃなく**
n-1で割る必要性があった。なぜか？

理由：母分散の推定には母平均 μ が必要だが、
未知なので、代わりに標本平均を代入したから。

偏差には $\sum_{i=1}^{n}(X_i-\bar{X})=0$ なる関係があるので
最初のn-1個が決まると最後の1個の値が決まる。

従って、実質的にはn個の和じゃなくてn-1個の和として表現でき、n-1で割ったほうがバランスが良い。

標本平均の代入の操作により未知パラメタが一つ消えているため**自由度**が１つ減るなどと言う。
(i.i.dの標本がn個ある場合、nの自由度とする)

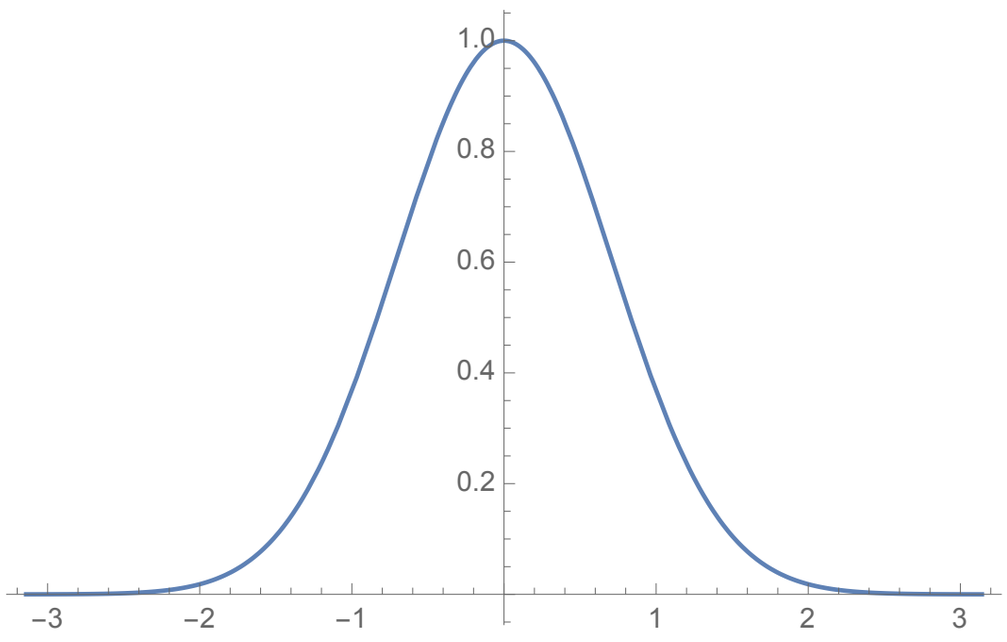
正規分布とその性質

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

平均μ、分散σ^2でiid

$$P\left(\frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leqslant \alpha\right) \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\alpha} \exp\left(-\frac{x^2}{2}\right) dx$$

中心極限定理により、独立な多数の因子の和として表される確率変数は正規分布に従う。



このことにより正規分布は統計学や自然科学、社会科学の様々な場面で複雑な現象を簡単に表すモデルとして用いられている。

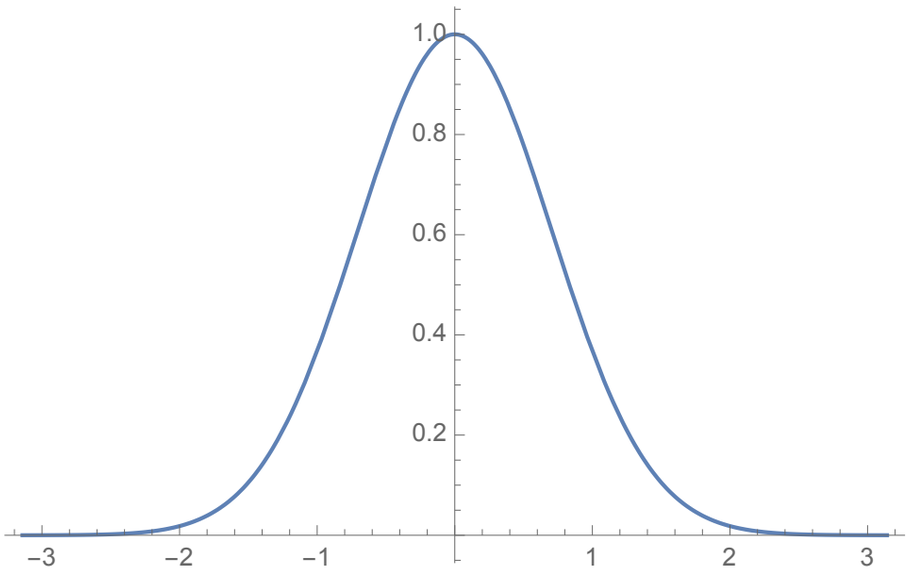
たとえば実験における測定の誤差は正規分布に従って分布すると仮定され、不確かさの評価が計算されている。

正規分布　(ガウス分布)

normal distribution　Gaussian distribution

基本：この関数の形に比例する確率密度

$$f(x) := \exp(-x^2) = e^{-x^2}$$



“釣鐘型”とか”ベル曲線(ベルカーブ)”
とか言われる。

(…が、どっちかっつうと山型？？)

確率密度なので積分したら1にする
ため、全面積を計算しておく

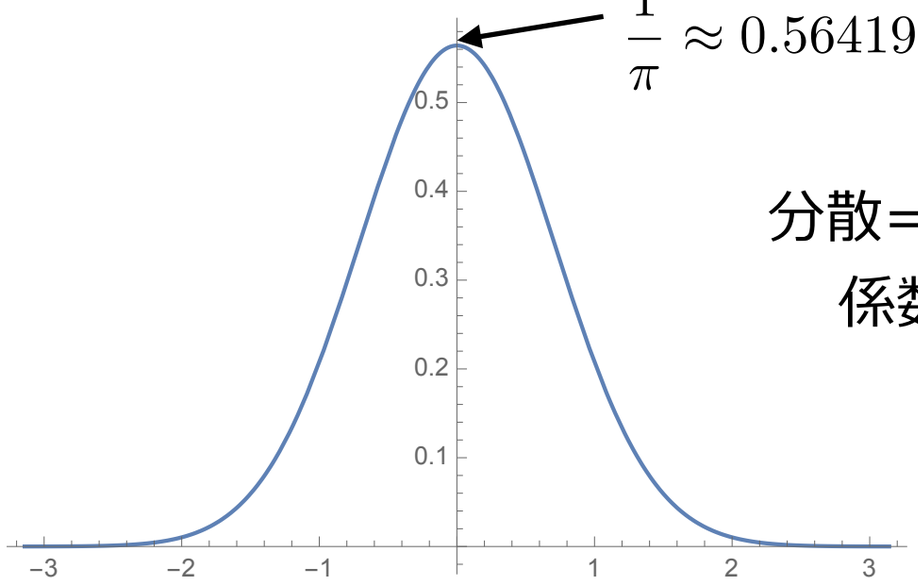
$$\int_{-\infty}^{\infty} \exp(-x^2)dx = \sqrt{\pi}$$

従って、この値で割って正規化した
以下の関数形は密度関数！

$$p(x) := \frac{1}{\sqrt{\pi}} \exp(-x^2)$$

ただしこのままだと分散が1/2で不格好!?

$$p(x) := \frac{1}{\sqrt{\pi}} \exp(-x^2)$$

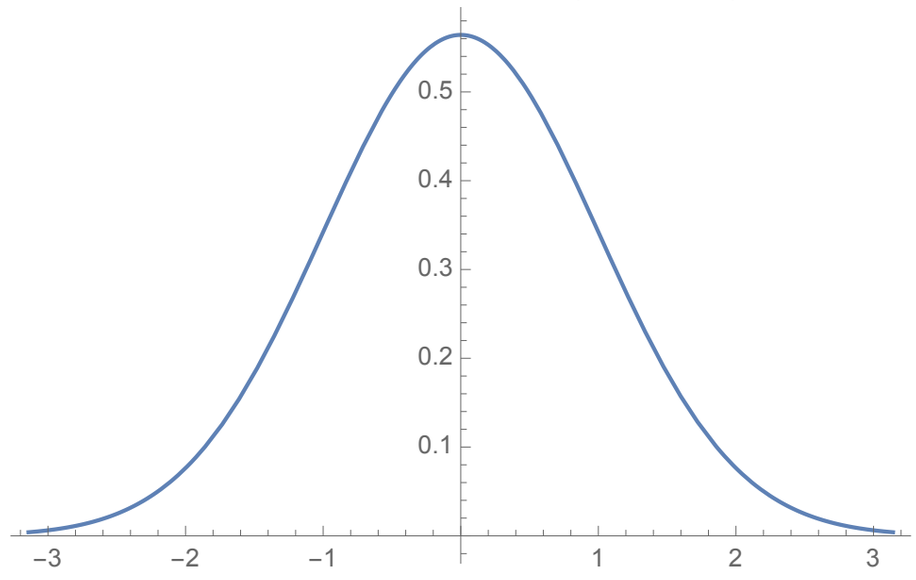


分散=1になるよう
係数をいじる

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}x^2\right) = \sqrt{2\pi} \quad \text{なので}$$

標準正規分布 $N(0,1)$

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



平均 0
分散 1

これが一般の単変量の正規分布を作るにも、多変量の正規分布を作るにも基本となる！！

平均(期待値)

$$\mathbb{E}\{x\} = \int_{-\infty}^{\infty} xp(x)dx = 0$$

分散

$$\mathbb{E}\{(x - \mathbb{E}\{x\})^2\} = \int_{-\infty}^{\infty} x^2 p(x)dx = \frac{1}{2}$$

分散=1/2

蛇足の復習）正規化項のガウス積分について

$$\int_{-\infty}^{\infty} \exp(-x^2)dx = \sqrt{\pi}$$

←――― この定積分は見た目ほど自明ではない

証明の例)　$I = \int_{-\infty}^{\infty} \exp(-x^2)dx$　と置いて、I^2　を計算してみると

$$I^2 = \left(\int_{-\infty}^{\infty} \exp(-x^2)dx\right)\left(\int_{-\infty}^{\infty} \exp(-y^2)dy\right)$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(-x^2-y^2)dxdy$$　↘ 極座標変換

$$= \int_{0}^{\infty}\int_{0}^{2\pi} \exp(-r^2)rd\theta dr = 2\pi\int_{0}^{\infty} \exp(-r^2)rdr = \pi$$

※以上により　$a > 0$　なら　$\int_{\infty}^{\infty} \exp\left(-(ax+b)^2\right) = \frac{\sqrt{\pi}}{a}$　も分かる！

慣れてください！(この形にギョっとしない)

$$Z := \frac{X - \mu}{\sqrt{\sigma^2}}$$

平均を引いて
標準偏差で割っただけ

X

平均μ,分散σ^2
の確率変数X

平均引いて標準偏差で割る

標準化

$$Z := \frac{X - \mu}{\sigma}$$

平均0,分散1
の確率変数Z

$$\mathbb{E}\{Z\} = \frac{\mathbb{E}\{X\} - \mu}{\sigma} = 0$$

$$\mathbb{E}\{(Z - \mathbb{E}\{Z\})^2\} = \mathbb{E}\{Z^2\}$$

$$= \frac{\mathbb{E}\{(X - \mu)^2\}}{\sigma^2} = 1$$

標準正規分布 → 平均μ 分散σ^2の正規分布

一般に平均μ,分散σ^2の確率変数Xを
平均0,分散1の確率変数Zに変換するには
以下の「標準化」をかます

積分値＝1にするための正規化定数は

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

確率変数の標準化

$$Z := \frac{X-\mu}{\sigma}$$

正規分布の密度関数

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

$$\mathbb{E}\{Z\} = \frac{\mathbb{E}\{X\}-\mu}{\sigma} = 0$$

$$\mathbb{E}\{(Z-\mathbb{E}\{Z\})^2\} = \mathbb{E}\{Z^2\}$$

$$= \frac{\mathbb{E}\{(X-\mu)^2\}}{\sigma^2} = 1$$

標準化　$Z \sim N(0,1)$

$$\Leftrightarrow X \sim N(\mu, \sigma^2)$$

基本：この関数の形に比例する確率密度

$$f(x) := \exp(-x^2) = e^{-x^2}$$

正規分布まとめ

❶

$$X \sim N(\mu, \sigma^2) \Rightarrow \frac{X - \mu}{\sqrt{\sigma^2}} \sim N(0, 1)$$

❷

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2) \Rightarrow \bar{X} = \frac{X_1 + \cdots + X_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

❸

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1)$$

← ❶と❷から言える❸が大事

❸

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

← ❶と❷から言える❸が大事

中心極限定理により、独立な多数の因子の和として表される確率変数は正規分布に従う。

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

平均μ、分散σ^2でiid

$$P\left(\frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leqslant \alpha\right) \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\alpha} \exp\left(-\frac{x^2}{2}\right) dx$$

$$\Leftrightarrow \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$