

3 章 データの統計的解析

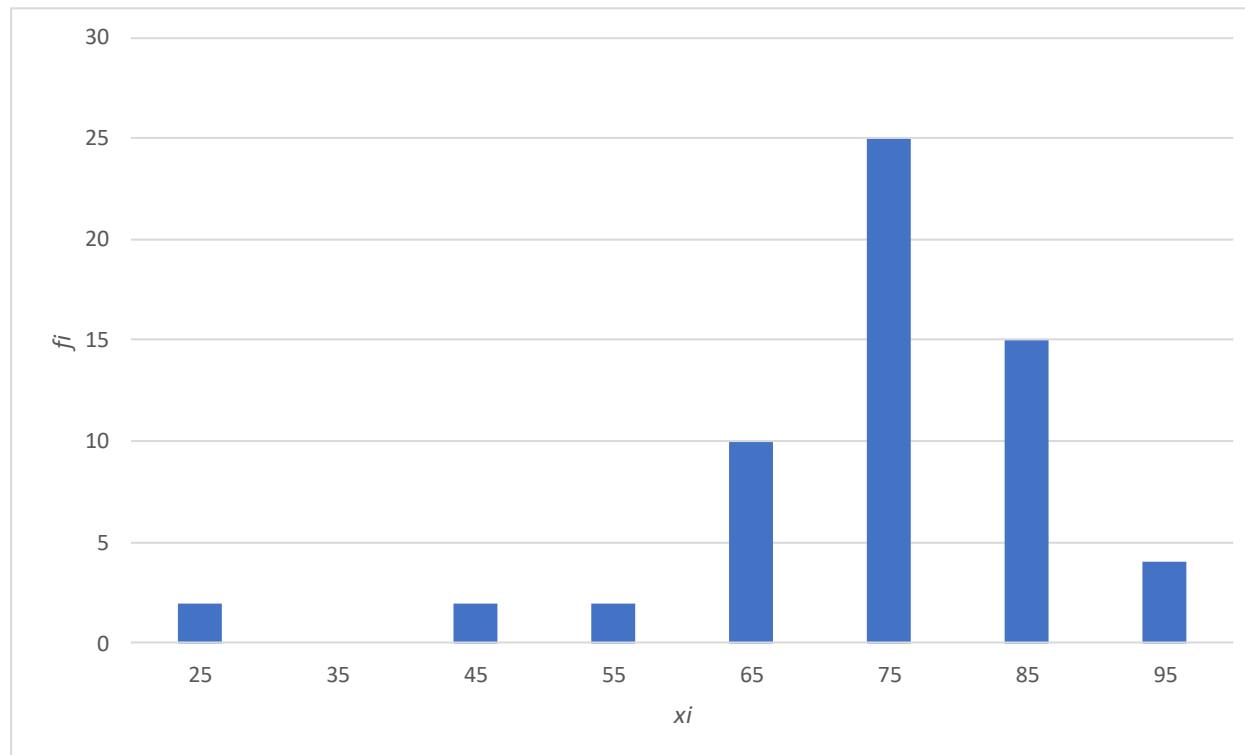
3.1 統計データの処理

例題：60 人の成績データ（10 点刻みに集計）

i	$x_{\min}$ (区間最低点数)	$x_{\max}$ (区間最大点数)	x_i (区間平均)	f_i (区間内人数)	$\sum_{j=1}^i f_j$ (累積人数)	$\frac{f_i}{n}$ (各区間人数/ 全人数)	$\frac{\sum_{j=1}^i f_j}{n}$ (累積人数/ 全人数)
1	20	30	25	2	2	0.033333	0.03333333
2	30	40	35	0	2	0	0.03333333
3	40	50	45	2	4	0.033333	0.06666667
4	50	60	55	2	6	0.033333	0.1
5	60	70	65	10	16	0.166667	0.26666667
6	70	80	75	25	41	0.416667	0.68333333
7	80	90	85	15	56	0.25	0.93333333
8	90	100	95	4	60	0.066667	1

クラスの成績を $i=1,2,\dots,8$ ($m=8$) の点数区間に分けてある。各区間 i は $x_{\min}$ より大きく $x_{\max}$ 以下であり、この区間の点数を取った学生の総数を f_i としてまとめている。

ヒストグラム



平均値 (mean)

$$\bar{X} = \frac{1}{n} \sum_{i=1}^m x_i f_i$$

$$\bar{X} = \frac{1}{60} (25 \times 2 + 35 \times 0 + 45 \times 2 + 55 \times 2 + 65 \times 10 + 75 \times 25 + 85 \times 15 + 95 \times 4) = 73.83 \quad (\text{点})$$

(注意：この平均値は 10 点刻みで分割したときの「近似」的な平均である)

分散 (variance)

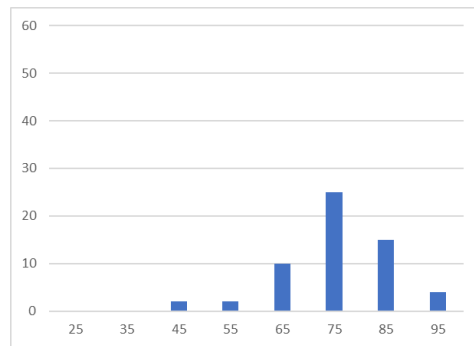
$$s^2 = \frac{1}{n} \sum_{i=1}^m (x_i - \bar{X})^2 f_i$$

分布の広がりを表す。

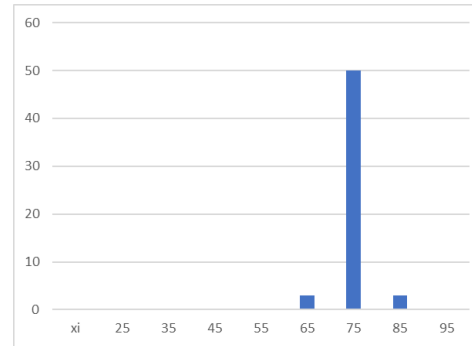
$$\begin{aligned} s^2 &= \frac{1}{60} \left\{ \begin{aligned} &(73.83 - 25)^2 \times 2 + (73.83 - 35)^2 \times 0 + (73.83 - 45)^2 \times 2 + (73.83 - 55)^2 \times 2 \\ &+ (73.83 - 65)^2 \times 10 + (73.83 - 75)^2 \times 25 + (73.83 - 85)^2 \times 15 \\ &+ (73.83 - 95)^2 \times 4 \end{aligned} \right\} \\ &= 601.7 \quad (\text{点}^2) \end{aligned}$$

標準偏差(standard deviation): $s = \sqrt{s^2}$

$$s = \sqrt{601.5} = 24.52 \quad (\text{点})$$



s が大きいクラス



s が小さいクラス

Figure 1 標準偏差とヒストグラムの関係

偏差値 (Standard score) : $T_i = \frac{10(x_i - \bar{X})}{s} + 50 \quad (1)$

ちょうど平均点を取る学生 ($x_i = \bar{X}$) の偏差値が 50 になる。

例：95 点の学生の偏差値は

$$T_i = \frac{10 \times (95 - 73.83)}{24.52} + 50 = 58.64$$

確率

変数 X が x_i になる ($X = x_i$) 確率は、

$$P\{X = x_i\}$$

または記号 p_i を使って

$$p_i = P\{X = x_i\} \quad (*)$$

次に、変数 X が x_i 以下になる ($X \leq x_i$) 確率は

$$P\{X \leq x_i\} \quad (**)$$

ここで、(*)を $i-1, i-2, \dots$ について書くと

$$p_{i-1} = P\{X = x_{i-1}\}$$

$$p_{i-2} = P\{X = x_{i-2}\}$$

$\dots$

$$p_1 = P\{X = x_1\}$$

従って、(**)は

$$P\{X \leq x_i\} = P\{X = x_i\} + P\{X = x_{i-1}\} + P\{X = x_{i-2}\} + \dots + P\{X = x_1\}$$

$$= \sum_{j=1}^i P\{X = x_j\} = \sum_{j=1}^i p_j$$

と表せる。

確率とヒストグラム

ヒストグラムの中から何か実現値 $\mathbf{X}$ を選んだとする。そして、ランダムに x_i をピックアップする。この x_i が $\mathbf{X}$ に等しい確率は

$$P\{X = x_i\} = p_i = \frac{f_i}{n}$$

と近似することが出来る（例：クラスからランダムで選んだ誰かの点数が 90 点より高く 100 点以下である確率）

上式を使うと、平均 μ は以下のように書き換えられる。

$$\mu = \frac{1}{n} \sum_{i=1}^m x_i f_i = \sum_{i=1}^m x_i \frac{f_i}{n} = \sum_{i=1}^m x_i \frac{f_i}{n} = \sum_{i=1}^m x_i p_i$$

この平均 μ は期待値(expected value)と呼ばれることもある。

$$\mu = E[X] \quad (*)$$

同様に分散 σ^2 は

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^m (x_i - \mu)^2 f_i = \sum_{i=1}^m (x_i - \mu)^2 \frac{f_i}{n} = \sum_{i=1}^m (x_i - \mu)^2 p_i$$

上の式は、(*)の x_i を $(x_i - \mu)^2$ に置き換えたものと見なすことが出来るので、

$$\sigma^2 = E\left[(x_i - \mu)^2\right]$$

と表すこともある。

確率分布関数

確率分布関数(probability distribution function) $F(x)$ を以下のように定義する。

$$F(x_i) = P\{X \leq x_i\} = \sum_{j=1}^i p_j$$

この関数は以下の性質を必ず持つ

$$F(-\infty) = P\{X \leq -\infty\} = 0$$

$$F(\infty) = P\{X \leq \infty\} = 1$$

確率変数 X が x_j と x_i の間にある確率は以下のように求めることが出来る。

$$\begin{aligned} P\{x_j < X \leq x_i\} &= P\{X \leq x_i\} - P\{X \leq x_j\} \\ &= F(x_i) - F(x_j) \end{aligned}$$

確立と確率分布関数

x_{i-1} と x_i の間隔を Δx とする。すなわち、

$$x_{i-1} + \Delta x = x_i$$

これを使うと

$$\begin{aligned} P\{x_{i-1} < X \leq x_i\} &= P\{x_i - \Delta x < X \leq x_i\} = P\{X < x_i\} - P\{X \leq x_i - \Delta x\} \\ &= F(x_i) - F(x_i - \Delta x) \quad (*) \end{aligned}$$

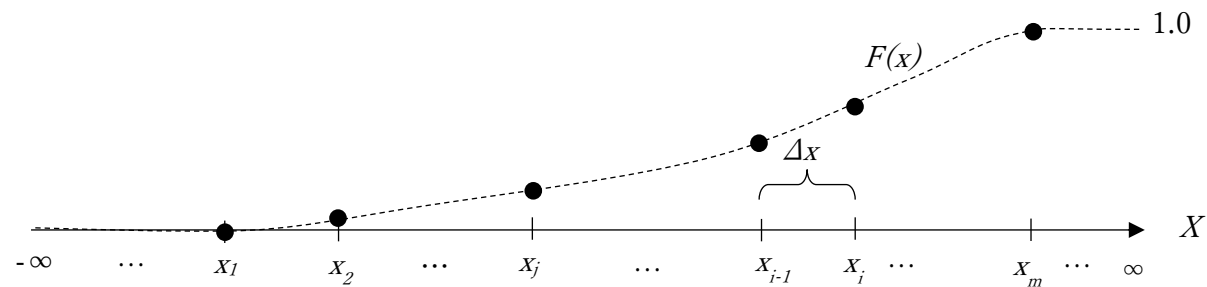
また、 $x_{i-1} < X \leq x_i$ に入る X の値は、 x_i だけであり、 x_{i-1} は含まれていないことに注意する（不等号が $<$ か $\leq$ に注意）。

$$P\{x_{i-1} < X \leq x_i\} = P\{x = x_i\} = p_i \quad (**)$$

この(*)と(**)より

$$p_i = F(x_i) - F(x_i - \Delta x)$$

この式は後に利用します。



二項分布と正規分布

30%の確率で表が出るコインを2回投げた時、表が x 回出る確率 $P\{X = x\}$ を考える。

$x=0$ のとき（一度も表が出ない）

$$P\{X = 0\} = (1-0.3) \times (1-0.3) = 0.7^2 = 0.49$$

$x=1$ のとき（一回目に投げた時に表が出るか、二回目に投げた時に表が出るか）

$$P\{X = 1\} = 0.3 \times 0.7 + 0.7 \times 0.3 = 2 \times 0.3 \times 0.7 = 0.42 \quad (*)$$

$x=2$ のとき（二度とも表が出る）

$$P\{X = 2\} = 0.3 \times 0.3 = 0.09$$

これを一般化する。(*)において係数2が出てきた理由は、コインを2回投げた時に一度でも表が出る全ての組み合わせの数

$${}_2C_1 = 2$$

従って、確率 p で表が出るコインを n 回投げた時に表が x 回出る確率は、以下のように表される。

$$P\{X = x\} = {}_nC_x p^x (1-p)^{n-x}$$

上野式を二項分布（binomial distribution）と呼ぶ。

また、 ${}_nC_x$ を以下の様に表現することもある。

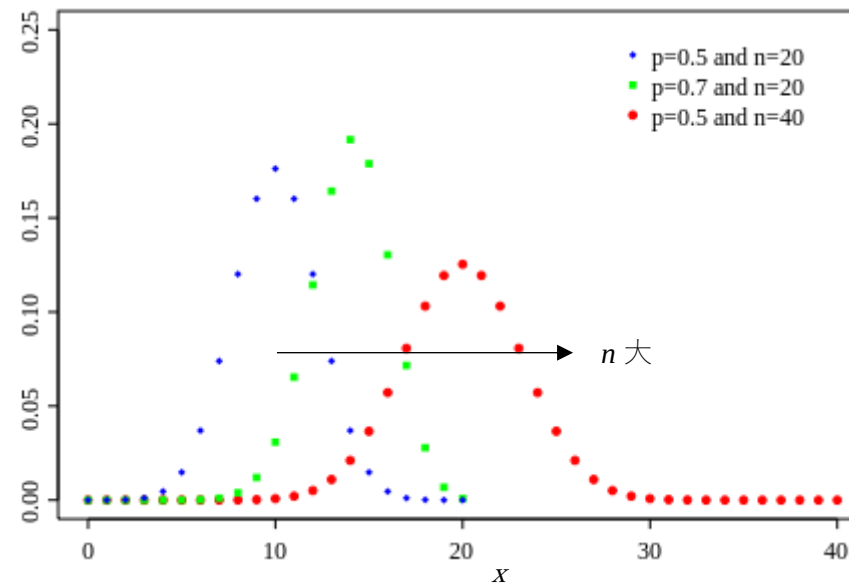
$${}_nC_x = \frac{n!}{(n-x)!x!} = \binom{n}{x}$$

この表記を使うと

$$P\{X = x\} = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x} = \binom{n}{x} p^x (1-p)^{n-x}$$

二項分布のプロット

n が大きくなると釣鐘型の分布に近づいていくことがわかる。



https://en.wikipedia.org/wiki/Binomial_distribution より転載。

二項分布と正規分布

n が十分大きい時、二項分布は以下の式に近づくことが知られている。

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

これは正規分布(normal distribution)またはガウス分布(Gaussian distribution)と呼ばれる。

ここで σ および μ は以下のように求められることが知られている。

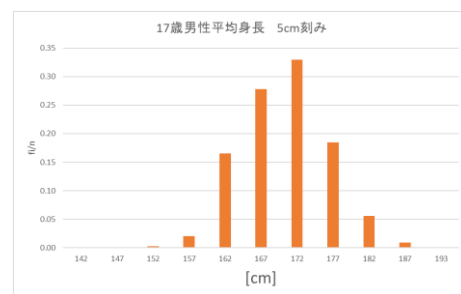
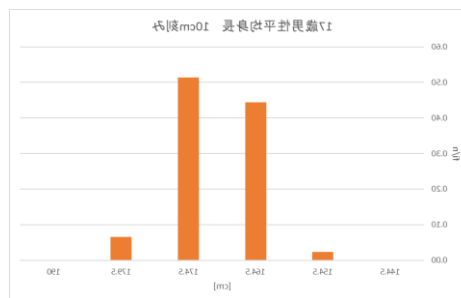
$$\begin{aligned}\mu &= np \\ \sigma &= \sqrt{np(1-p)}\end{aligned}$$

特に $\mu=0$ 、 $\sigma=1$ の正規分布を標準正規分布(Standard normal distribution)と呼ぶ。

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

離散確率変数と連続確率変数

例：日本人の身長（17 歳男性）をヒストグラムにするとどうなるか。



だんだん連続の関数に近くなっていく

平均と分散の計算

平均

$$\mu = \sum_{i=1}^m x_i p_i = \sum_{i=1}^m x_i \{F(x_i) - F(x_i - \Delta x)\}$$

これを変形する

$$\mu = \sum_{i=1}^m x_i \{F(x_i) - F(x_i - \Delta x)\} = \sum_{i=1}^m x_i \frac{F(x_i) - F(x_i - \Delta x)}{\Delta x} \Delta x$$

ここで $\Delta x \rightarrow 0$ の極限を考えると

$$\mu = \lim_{\Delta x \rightarrow 0} \sum_{i=1}^m x_i \frac{F(x_i) - F(x_i - \Delta x)}{\Delta x} \Delta x = \int_{x_1}^{x_m} x \frac{dF(x)}{dx} dx$$

ここで、積分の中にある導関数を確率密度関数（probability density function） $f(x)$ とする。

$$\frac{dF(x)}{dx} = f(x)$$

と定義する。これを使うと平均は

$$\mu = \int_{x_1}^{x_m} x f(x) dx$$

平均と分散の計算（続き）

導出した平均は

$$\mu = \int_{x_1}^{x_m} x f(x) dx$$

しかし通常、平均は上限を $+\infty$ 、下限を $-\infty$ として

$$\mu = \int_{-\infty}^{\infty} x f(x) dx$$

と定義することが多い。これは、 $x < x_1$ および $x > x_m$ において $f(x) = 0$ と仮定したものである。

平均（期待値）の別の表記として、 $E[X]$ を使い

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

この表記を使うと、分散は

$$\sigma^2 = E\left[(X - \mu)^2\right] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

確率密度関数の性質

確率密度関数 $f(x)$ を使っていろいろな確率を表してみる。

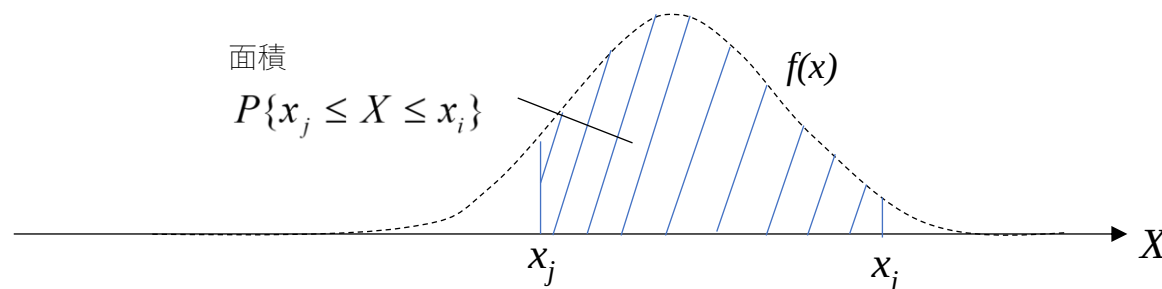
$$P\{X \leq x_i\} = F(x_i) = F(x_i) - \underbrace{F(-\infty)}_{=0} = \int_{-\infty}^{x_i} \frac{dF(x)}{dx} dx = \int_{-\infty}^{x_i} f(x) dx$$

これを使うと、

$$P\{x_j \leq X \leq x_i\} = P\{X \leq x_i\} - P\{X \leq x_j\}$$

$$= \int_{-\infty}^{x_i} f(x) dx - \int_{-\infty}^{x_j} f(x) dx$$

$$= \int_{x_j}^{x_i} f(x) dx$$

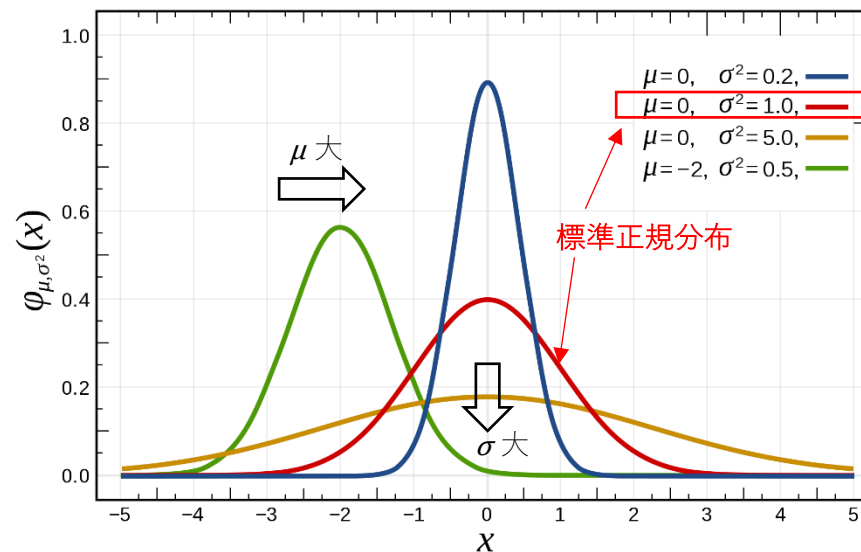


正規分布：確率密度関数の一種

正規分布も確率密度関数の一つである。

$$\text{確率密度関数： } f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

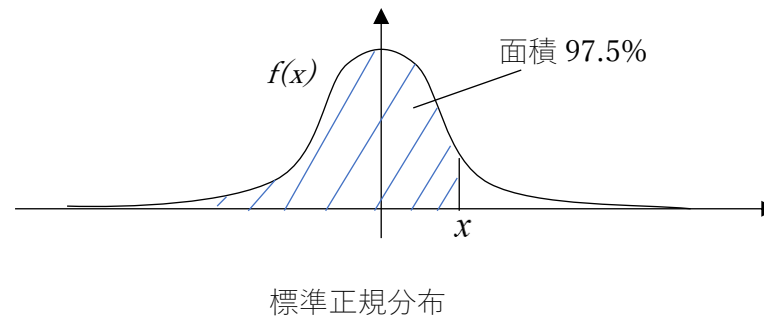
$$\text{確率分布関数： } F(x) = P\{X \leq x_i\} = \int_{-\infty}^{x_i} f(x)dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x_i} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$



特に標準正規分布（ $\mu=0$, $\sigma=1$ ）は、教科書巻末の付録 1 に値が載っている。

https://en.wikipedia.org/wiki/Normal_distribution

例題 1：確率密度関数 $f(x)$ が標準正規分布に従うとき、 $P\{X \leq x\} = 0.975$ となる x を求めよ。



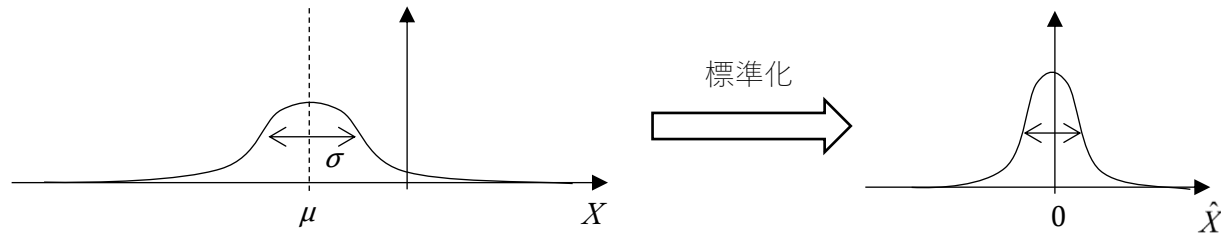
解答：教科書付録 1 で $\Phi(x) = 0.975$ に一番近い値は、 $x = 1.96$ である。従って、 $x = 1.96$ 。

(解答終)

確率変数 X が正規分布に従うが、標準正規分布ではない場合 ($\mu \neq 0, \sigma \neq 1$)。

標準化する

$$\hat{X} = \frac{X - \mu}{\sigma}$$



例 2：確率変数 X が $\mu=5, \sigma=2$ の正規分布に従うとする。 $P\{X \leq 3\}$ を求めよ。

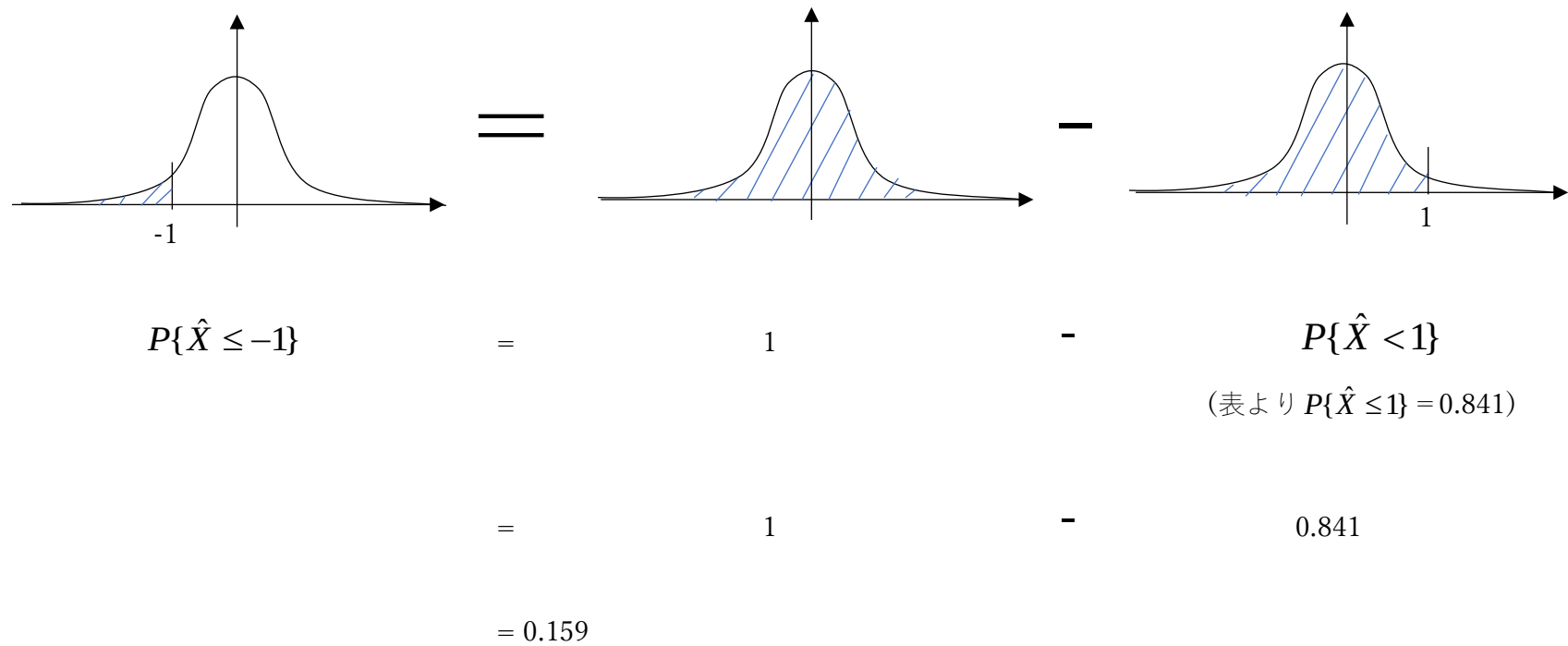
解答：標準化を行う。 $x=3$ のときを考えると、

$$\hat{x} = \frac{x - \mu}{\sigma} = \frac{3 - 5}{2} = -1$$

すなわち、

$$P\{X \leq 3\} = P\{\hat{X} \leq -1\}$$

正規分布は左右対称であることを利用する



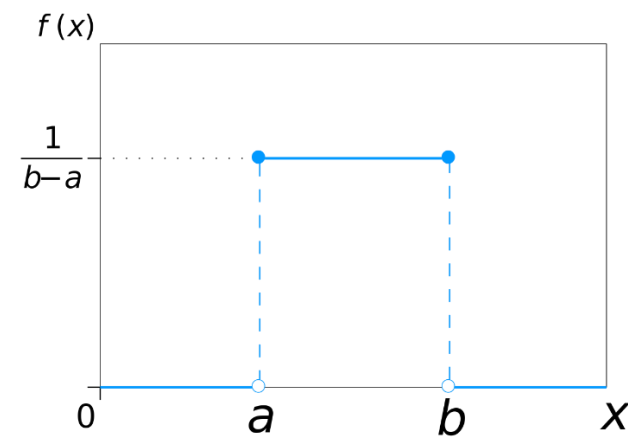
(解答終)

いろいろな確率分布

1. 一様分布 (uniform distribution)

確率変数 x が範囲 $a \leq x \leq b$ において全て一定の確率を取る。

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & x < a, x > b \end{cases}$$



確率分布関数の性質

$$F(\infty) = \int_{-\infty}^{\infty} f(x) dx = 1$$

は簡単に確かめられる

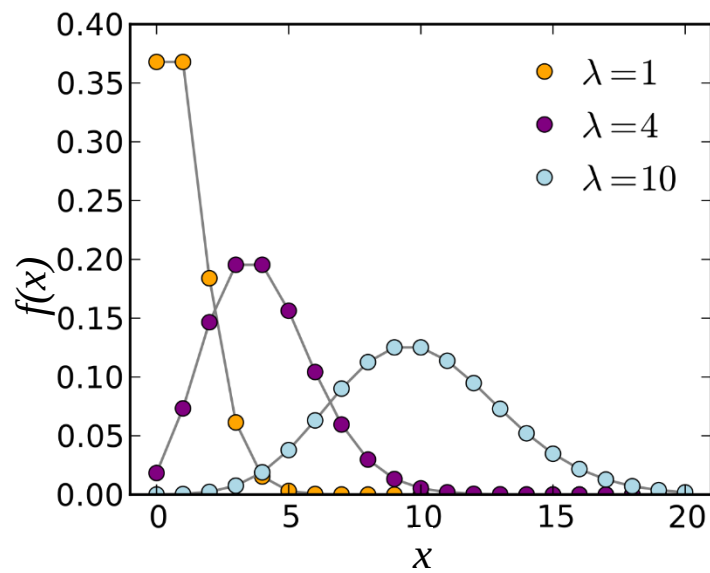
[https://en.wikipedia.org/wiki/Uniform_distribution_\(continuous\)](https://en.wikipedia.org/wiki/Uniform_distribution_(continuous))

2. ポアソン分布 (Poisson distribution)

離散（整数）確率変数 x を考える。一定時間以内に平均 λ 回起こる現象が x 回起きる確率を表すときに用いられる。

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

例：年間に起こる地震の平均が 10 回（ $\lambda=10$ ）のとき、今年に起こる地震の回数は？

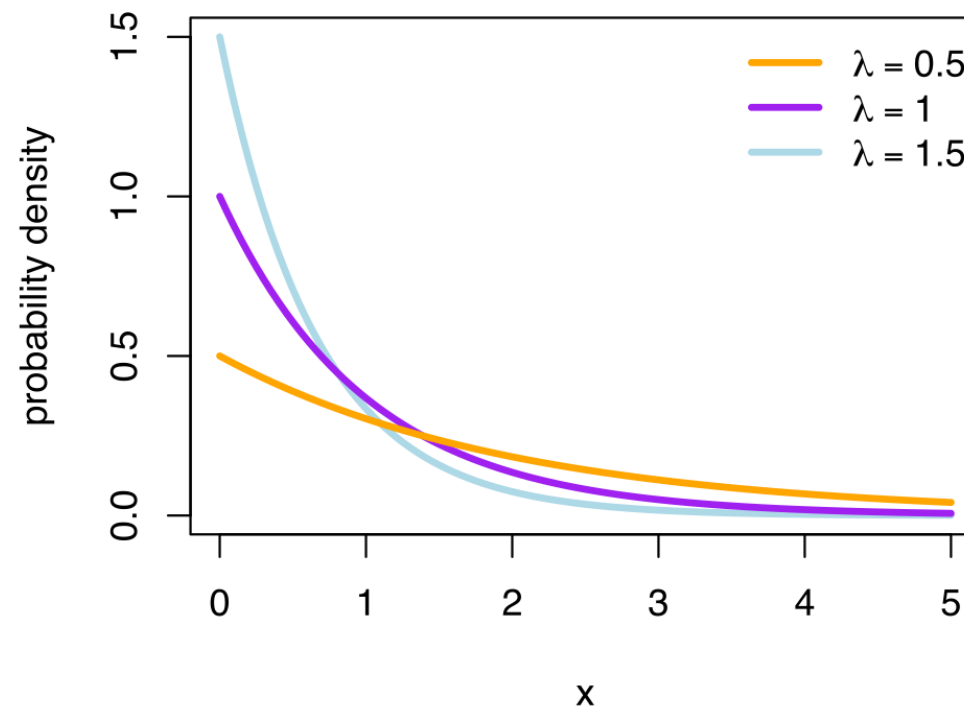


https://en.wikipedia.org/wiki/Poisson_distribution

3. 指数分布

$$f(x) = \lambda e^{-\lambda x}$$

ただし $x \geq 0$ でのみ定義する。



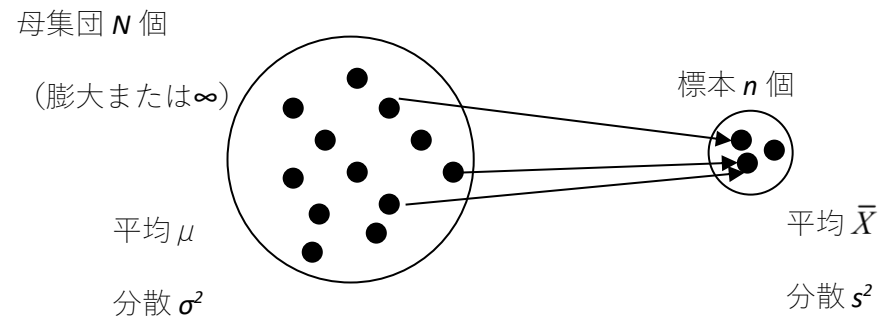
https://en.wikipedia.org/wiki/Exponential_distribution

母集団と標本

統計を取りたい量が膨大な場合、標本(sample)を取ることで母集団(population)を推定する。

例題 3. 1 あるシステムが故障するための時間を測定する実験を 15 回行った結果を以下に示す。
37, 3, 13, 42, 23, 31, 10, 16, 30, 29, 20, 19, 25, 28, 33 (h)

この実験を多く繰り返すと更に多くのデータが得られる。



母集団と標本：平均と分散

標本 $x_1, x_2, \dots, x_n$ が得られたとする。

標本の平均が計算できる

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i \quad \rightarrow \quad \text{これを使って母集団の平均を推定}$$

標本の分散

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2 \quad \rightarrow \quad \text{これを使って母集団の分散を推定（？）}$$

	母集団	標本
平均	$\mu = \frac{1}{N} \sum_{i=1}^N x_i$	$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$
分散	$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$	$s^2 = \frac{1}{\textcolor{red}{n}-1} \sum_{i=1}^n (x_i - \bar{X})^2$

母集団の分散 σ^2 を推定する標本の分散 s^2 は、分母に $\textcolor{red}{n}$ ではなく $\textcolor{red}{n}-1$ となることに注意する。理由は割愛する（標本を新しく取りなおすたびに変動しうることによる）。

例題 3. 1 あるシステムが故障するための時間を測定する実験を 15 回行った結果を以下に示す。

$$37, 3, 13, 42, 23, 31, 10, 16, 30, 29, 20, 19, 25, 28, 33 \quad (\text{h})$$

故障時間 $x(\text{h})$ が正規分布に従うと仮定する。

$$\text{平均} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{15} (37 + 3 + 13 + 42 + \dots) = 23.9 \quad (\text{h})$$

分散

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2 = \frac{1}{15-1} \left\{ (37-23.9)^2 + (3-23.9)^2 + (13-23.9)^2 + (42-23.9)^2 + \dots \right\} = 112 \quad (\text{h}^2)$$

これより

$$s = \sqrt{112} = 10.6 \quad (\text{h})$$

これらを使って母集団の平均 $\mu \approx \bar{X}$ 、分散 $\sigma \approx s$ と推定する。

故障時間 $x(\text{h})$ は正規分布に従うので、

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

推定された平均 $\mu \approx \bar{X}$ 、分散 $\sigma \approx s$ を使って、 $f(x) = \frac{1}{\sqrt{2\pi \times 10.6^2}} \exp\left(-\frac{(x-23.9)^2}{2 \times 10.6^2}\right)$ と推定される。

これはあくまで推定値。実験回数を 15 回より更に増やしていけば、標本は母集団に近づいていくので、真の確率密度分布を求められる。