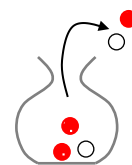


確率分布

§1 確率変数と確率分布

『赤玉 3 個、白玉 2 個が入っている袋から、玉を同時に 2 個取り出す』試行において、白玉を X 個取り出す確率を考える。このとき、 X のとり得る値は $X = 0, 1, 2$ であり、その各値における確率 P は右の表のようになる。



この例のように、試行の結果によってある値 X が決まり、その X に対して確率が定まるとき、その変数 X を**確率変数**という。また、確率変数と確率の対応関係を表す表を**確率分布**という。

X	0	1	2	計
P	$\frac{3}{10}$	$\frac{3}{5}$	$\frac{1}{10}$	1

一般的に、ある試行において、確率変数 X のとる値が $x_1, x_2, x_3, \dots, x_n$ であるとき、 $X = x_k$ となる確率を $P(X = x_k)$ と表す。また、 X の値が a 以上 b 以下である確率を $P(a \leq X \leq b)$ と表す。

例えば上の例だと、

$$P(X = 1) = \frac{3}{5}, \quad P(X = 2) = \frac{1}{10}$$

$$P(1 \leq X \leq 2) = \frac{3}{5} + \frac{1}{10} = \frac{7}{10}$$

となる。

また、 $P(X = x_k) = p_k$ とすると確率分布は右のようになる。

このとき確率変数 X はこの分布に**従う**といい、 $p_1, p_2, \dots, p_n$ は

$$p_1 \geq 0, p_2 \geq 0, \dots, p_n \geq 0$$

$$p_1 + p_2 + \dots + p_n = 1$$

を満たす。

X	x_1	x_2	$\dots$	x_n	計
P	p_1	p_2	$\dots$	p_n	1

○確率変数の期待値

あるくじ引きにおいて、くじの本数と賞金の関係は右の表のようになっているとする。賞金を X とするとき、 X の確率分布は以下のようにになる。

X	10000	5000	1000	100	計
P	$\frac{5}{100}$	$\frac{15}{100}$	$\frac{25}{100}$	$\frac{55}{100}$	1

	賞金	本数
1等	10,000	5
2等	5,000	15
3等	1,000	25
4等	100	55

このとき、 X の期待値(expectation)は次のように計算することができる。

$$10000 \times \frac{5}{100} + 5000 \times \frac{15}{100} + 1000 \times \frac{25}{100} + 100 \times \frac{55}{100} = 1555$$

つまり、期待値は『**(確率変数)×(確率)の和**』で求められる。

一般的に、確率変数 X が右の分布に従うとき、

$$x_1 p_1 + x_2 p_2 + \dots + x_n p_n = \sum_{k=1}^n x_k p_k$$

を確率変数 X の**平均**または**期待値**といい $E(X)$ と表す。

X	x_1	x_2	$\dots$	x_n	計
P	p_1	p_2	$\dots$	p_n	1

確率変数の期待値

$P(X = x_k) = p_k$ ($k = 1, 2, \dots, n$) のとき、期待値 $E(X)$ は次のようになる。

$$E(X) = x_1 p_1 + x_2 p_2 + \dots + x_n p_n = \sum_{k=1}^n x_k p_k$$

○確率変数の分散と標準偏差

期待値を考えたときと同様に、右表のようなくじ引きについて考える。

ここでは、このくじ引きの分散(variance)、標準偏差(standard deviation)を考える。

	賞金	本数
1等	10,000	5
2等	5,000	15
3等	1,000	25
4等	100	55

このくじ引きの期待値(平均)は 1555 (円) なので、分散は以下のようになる。

$$\frac{(10000 - 1555)^2 \times 5 + (5000 - 1555)^2 \times 15 + (1000 - 1555)^2 \times 25 + (100 - 1555)^2 \times 55}{100} = 6587475$$

この式の左辺は次のように変形することができる。

$$(10000 - 1555)^2 \times \frac{5}{100} + (5000 - 1555)^2 \times \frac{15}{100} + (1000 - 1555)^2 \times \frac{25}{100} + (100 - 1555)^2 \times \frac{55}{100}$$

この式において、 $\frac{\quad}{100}$ 部分はそれぞれ 1 等, 2 等, 3 等, 4 等が当たる確率となっている。

つまり、分散は『(偏差)² × (確率) の和』で求められることが分かる。

なお、標準偏差は $\sqrt{(\text{分散})}$ なので、 $\sqrt{6587475} = 2566.607 \dots$ (円) となる。

一般的に、確率変数 X が右の分布に従い、期待値を $E(X) = m$ とするとき、

$$(x_1 - m)^2 p_1 + (x_2 - m)^2 p_2 + \dots + (x_n - m)^2 p_n = \sum_{k=1}^n (x_k - m)^2 p_k$$

X	x_1	x_2	$\dots$	x_n	計
P	p_1	p_2	$\dots$	p_n	1

を確率変数 X の**分散**といい $V(X)$ と表す。

また、 $\sqrt{V(X)}$ を確率変数 X の**標準偏差**といい $\sigma(X)$ と表す(σ は s に相当するギリシャ文字で「シグマ」と読む)。

確率変数の分散・標準偏差①

$P(X = x_k) = p_k$ ($k = 1, 2, \dots, n$), $E(X) = m$ のとき、分散 $V(X)$, 標準偏差 $\sigma(X)$ は次のようになる。

$$V(X) = (x_1 - m)^2 p_1 + (x_2 - m)^2 p_2 + \dots + (x_n - m)^2 p_n = \sum_{k=1}^n (x_k - m)^2 p_k$$

$$\sigma(X) = \sqrt{(x_1 - m)^2 p_1 + (x_2 - m)^2 p_2 + \dots + (x_n - m)^2 p_n} = \sqrt{\sum_{k=1}^n (x_k - m)^2 p_k}$$

X が確率変数なので、 $(X - m)^2$ も確率変数となり、以下の分布に従う。

$(X - m)^2$	$(x_1 - m)^2$	$(x_2 - m)^2$	$\dots$	$(x_n - m)^2$	計
P	p_1	p_2	$\dots$	p_n	1

これより、分散は確率変数 $(X - m)^2$ の期待値 $E((X - m)^2)$ と見ることができる。

なお、分散の式は以下のように変形することができる。

$$V(X) = \sum_{k=1}^n (x_k - m)^2 p_k = \sum_{k=1}^n (x_k^2 - 2mx_k + m^2) p_k = \sum_{k=1}^n x_k^2 p_k - 2m \sum_{k=1}^n x_k p_k + m^2 \sum_{k=1}^n p_k$$

$$\sum_{k=1}^n x_k p_k = m, \quad \sum_{k=1}^n p_k = 1 \text{ より,}$$

$$V(X) = \sum_{k=1}^n x_k^2 p_k - 2m \cdot m + m^2 \cdot 1 = \sum_{k=1}^n x_k^2 p_k - m^2$$

ここで、 X^2 は確率変数なので、 $\sum_{k=1}^n x_k^2 p_k$ は確率変数 X^2 の期待値と見ることができる。つまり、 $E(X^2)$ と表せるので、

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

と表すことができる。

確率変数 X^2 は以下の分布に従う。

X^2	x_1^2	x_2^2	$\cdots$	x_n^2	計
P	p_1	p_2	$\cdots$	p_n	1

確率変数の分散・標準偏差②

$$V(X) = E(X^2) - \{E(X)\}^2, \quad \sigma(X) = \sqrt{E(X^2) - \{E(X)\}^2}$$

○確率変数の変換

確率変数 X が右のような分布に従うとする。

このとき、 $aX + b$ も確率変数となり、以下の確率分布に従う。

X	x_1	x_2	$\cdots$	x_n	計
P	p_1	p_2	$\cdots$	p_n	1

$aX + b$	$ax_1 + b$	$ax_2 + b$	$\cdots$	$ax_n + b$	計
P	p_1	p_2	$\cdots$	p_n	1

これより、確率変数 $aX + b$ の平均、分散、標準偏差は次のようになる。

$$E(aX + b) = \sum_{k=1}^n (ax_k + b) p_k = a \sum_{k=1}^n x_k p_k + b \sum_{k=1}^n p_k = aE(X) + b$$

$E(X) = m$ とおくと、

$$V(aX + b) = \sum_{k=1}^n \{(ax_k + b) - (am + b)\}^2 p_k = a^2 \sum_{k=1}^n (x_k - m)^2 p_k = a^2 V(X)$$

$$\sigma(aX + b) = \sqrt{V(aX + b)} = \sqrt{a^2 V(X)} = |a| \sigma(X)$$

確率変数の変換

$$E(aX + b) = aE(X) + b, \quad V(aX + b) = a^2 V(X), \quad \sigma(aX + b) = |a| \sigma(X)$$

§ 2 確率変数の積と和

ここでは、複数の確率変数の和または積の期待値や分散がどのように表すことができるかを考えていく。

○同時確率分布

例1 赤玉2個、白玉4個が入っている袋から、まずAさんが1個取り出し、次にBさんが2個取り出す。ただし、取り出した玉は元に戻さないとする。このとき、次の確率を求めなさい。

- (1) Aさんが白玉を、Bさんが白玉1個、赤玉1個を取り出す確率。
- (2) Bさんが白玉1個を取り出す確率。

事象A, B, Cを次のように定める。

A : Aさんが白玉を取り出す B : Aさんが赤玉を取り出す

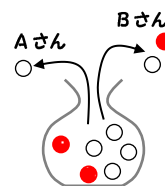
C : Bさんが白玉1個、赤玉1個を取り出す

- (1) 求める確率は

$$P(A \cap C) = P(A) \cdot P_A(C) = \frac{4}{6} \cdot \frac{2 \cdot 3}{5C_2} = \frac{2}{3} \cdot \frac{3}{5} = \frac{2}{5}$$

- (2) 求める確率は

$$P(C) = P(A \cap C) + P(B \cap C) = \frac{2}{5} + \frac{2}{6} \cdot \frac{1 \cdot 4}{5C_2} = \frac{8}{15}$$



Aさんが取り出す白玉の個数を X 、Bさんが取り出す白玉の個数を Y とすると、 X, Y は確率変数となり、各 X, Y における確率は右のような2次元の表になる。

このように、2つの確率変数 X, Y と確率の対応関係を表す表を**同時確率分布**という。なお、確率変数 X, Y のそれぞれの確率分布は右表の下端と右端の「計」の欄を見ることで、以下のようになっていることが分かる。

X	0	1	計
P	$\frac{1}{3}$	$\frac{2}{3}$	1

Y	0	1	2	計
P	$\frac{1}{15}$	$\frac{8}{15}$	$\frac{2}{5}$	1

$X \backslash Y$	0	1	2	計
0	0	$\frac{2}{15}$	$\frac{1}{5}$	$\frac{1}{3}$
1	$\frac{1}{15}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{2}{3}$
計	$\frac{1}{15}$	$\frac{8}{15}$	$\frac{2}{5}$	1

一般的に、ある試行において、確率変数 X のとり値が $x_1, x_2, \dots, x_m$ 、確率変数 Y のとり値が $y_1, y_2, \dots, y_n$ であるとき、

$$P(X = x_i, Y = y_j) = p_{ij}$$

$$P(X = x_i) = p_{i1} + p_{i2} + \dots + p_{in} = p_i$$

$$P(Y = y_j) = p_{1j} + p_{2j} + \dots + p_{mj} = q_j$$

と表すと、同時確率分布は右のようになる。

このとき p_i, q_j ($i = 1, 2, \dots, m, j = 1, 2, \dots, n$) は

$$p_1 + p_2 + \dots + p_m = q_1 + q_2 + \dots + q_n = 1$$

を満たす。

$X \backslash Y$	y_1	y_2	$\dots$	y_j	$\dots$	y_n	計
x_1	p_{11}	p_{12}	$\dots$	p_{1j}	$\dots$	p_{1n}	p_1
x_2	p_{21}	p_{22}	$\dots$	p_{2j}	$\dots$	p_{2n}	p_2
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_i	p_{i1}	p_{i2}	$\dots$	p_{ij}	$\dots$	p_{in}	p_i
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_m	p_{m1}	p_{m2}	$\dots$	p_{mj}	$\dots$	p_{mn}	p_m
計	q_1	q_2	$\dots$	q_j	$\dots$	q_n	1

○確率変数の和の期待値

2つの確率変数 X , Y に対して、和 $X + Y$ も確率変数となる。ここでは確率変数 $X + Y$ の期待値を考えよう。

例2 赤玉 2 個、白玉 4 個が入っている袋から、まず A さんが 1 個取り出し、次に B さんが 2 個取り出す。ただし、取り出した玉は元に戻さないとする。このとき、取り出された白玉の個数の和の期待値を求めなさい。

問題の設定は**例1**と同じです。

A さんが取り出す白玉の個数 X , B さんが取り出す白玉の個数を Y とすると、 X , Y の同時確率分布は右のようになる。

これより、 $X + Y$ の確率分布は以下のようになる。

$X + Y$	0	1	2	3	計
P	0	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{1}{5}$	1

これより期待値 $E(X + Y)$ は、次のようになる。

$$E(X + Y) = 0 \times 0 + 1 \times \frac{1}{5} + 2 \times \frac{3}{5} + 3 \times \frac{1}{5} = 2$$

$X \backslash Y$	0	1	2	計
0	0	$\frac{2}{15}$	$\frac{1}{5}$	$\frac{1}{3}$
1	$\frac{1}{15}$	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{2}{3}$
計	$\frac{1}{15}$	$\frac{8}{15}$	$\frac{2}{5}$	1

例2において、 X , Y の確率分布が右のようになるので、期待値 $E(X)$, $E(Y)$ は以下のようになる。

$$E(X) = 0 \times \frac{1}{3} + 1 \times \frac{2}{3} = \frac{2}{3}$$

$$E(Y) = 0 \times \frac{1}{15} + 1 \times \frac{8}{15} + 2 \times \frac{2}{5} = \frac{4}{3}$$

X	0	1	計
P	$\frac{1}{3}$	$\frac{2}{3}$	1

Y	0	1	2	計
P	$\frac{1}{15}$	$\frac{8}{15}$	$\frac{2}{5}$	1

これより、 $E(X + Y) = E(X) + E(Y)$ という関係式が成り立っていることが分かる。

一般的に、 $P(X = x_i, Y = y_j) = p_{ij}$ であるとき、 $P(X + Y = x_i + x_j) = p_{ij}$ なので、 $X + Y$ の期待値 $E(X + Y)$ は以下のように表すことができる。

$$\begin{aligned}
 E(X + Y) &= (x_1 + y_1)p_{11} + (x_1 + y_2)p_{12} + \cdots + (x_1 + y_n)p_{1n} \\
 &\quad + (x_2 + y_1)p_{21} + (x_2 + y_2)p_{22} + \cdots + (x_2 + y_n)p_{2n} \\
 &\quad + \cdots \\
 &\quad + (x_m + y_1)p_{m1} + (x_m + y_2)p_{m2} + \cdots + (x_m + y_n)p_{mn} \\
 &= x_1 \sum_{k=1}^n p_{1k} + x_2 \sum_{k=1}^n p_{2k} + \cdots + x_m \sum_{k=1}^n p_{mk} \\
 &\quad + y_1 \sum_{k=1}^m p_{k1} + y_2 \sum_{k=1}^m p_{k2} + \cdots + y_n \sum_{k=1}^m p_{kn} \\
 &= (x_1 p_{11} + x_2 p_{21} + \cdots + x_m p_{m1}) + (y_1 q_1 + y_2 q_2 + \cdots + y_n q_n) \\
 &= E(X) + E(Y)
 \end{aligned}$$

$X \backslash Y$	y_1	y_2	$\cdots$	y_n	計
x_1	p_{11}	p_{12}	$\cdots$	p_{1n}	p_1
x_2	p_{21}	p_{22}	$\cdots$	p_{2n}	p_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_m	p_{m1}	p_{m2}	$\cdots$	p_{mn}	p_m
計	q_1	q_2	$\cdots$	q_n	1

また、この等式と、確率変数の変換の等式を合わせると、
以下の関係が成り立っていることが分かる。

確率変数の変換

$$E(aX + b) = aE(X) + b$$

和の期待値

$$\begin{aligned} E(X + Y) &= E(X) + E(Y) \\ E(aX + bY) &= aE(X) + bE(Y) \end{aligned}$$

このことを利用すると、3つ以上の確率変数の和についても同様の関係が成り立つ。

$$E(X + Y + Z) = E((X + Y) + Z) = E(X + Y) + E(Z) = E(X) + E(Y) + E(Z)$$

例3 赤玉3個、白玉6個が入っている袋から、3個取り出す。取り出された赤玉の個数の期待値を求めなさい。

解①

取り出す赤玉の個数を X 個とすると、 $X = 0, 1, 2, 3$ となる。

$$P(X = 0) = \frac{{}_6C_3}{{}_9C_3} = \frac{5}{21}, \quad P(X = 1) = \frac{3 \cdot {}_6C_2}{{}_9C_3} = \frac{15}{28}$$

$$P(X = 2) = \frac{{}_3C_2 \cdot 6}{{}_9C_3} = \frac{3}{14}, \quad P(X = 3) = \frac{1}{{}_9C_3} = \frac{1}{84}$$

これより、 X の確率分布は右のようになる。

よって、期待値 $E(X)$ は、次のようになる。

$$E(X) = 0 \times \frac{5}{21} + 1 \times \frac{15}{28} + 2 \times \frac{3}{14} + 3 \times \frac{1}{84} = 1$$

X	0	1	2	3	計
P	$\frac{5}{21}$	$\frac{15}{28}$	$\frac{3}{14}$	$\frac{1}{84}$	1

袋から1個取り出すとき、赤玉を引く確率は $\frac{3}{9} = \frac{1}{3}$

袋から2個取り出すとき、2個目に赤玉を引く確率は $\frac{3}{9} \cdot \frac{2}{8} + \frac{6}{9} \cdot \frac{3}{8} = \frac{1}{3}$

このように、袋から1個ずつ順番に玉を取り出すとき、赤玉を取り出す確率は何回目であっても常に同じになります。これはくじ引きが引く順番で有利・不利がないのと同じです。このことと和の期待値の性質を利用すると、次のように求めることができます。

解②

3個の玉を1個ずつ順に3個取り出すと考える。

n 回目に取り出す赤玉の個数を X_n 個とすると、 $X_n = 0, 1$ となる。

このとき、 X_n の確率分布は右のようになるので、期待値 $E(X_n)$ は以下のようになる。

$$E(X_n) = 0 \cdot \frac{2}{3} + 1 \cdot \frac{1}{3} = \frac{1}{3}$$

取り出す赤玉の個数を X 個とすると、

$$E(X) = E(X_1 + X_2 + X_3) = E(X_1) + E(X_2) + E(X_3) = \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1$$

X_n	0	1	計
P	$\frac{2}{3}$	$\frac{1}{3}$	1

○独立な確率変数

例 4 赤玉 2 個, 白玉 4 個が入っている袋から, まず A さんが 1 個取り出し, 次に B さんが 2 個取り出す。ただし, 取り出した玉は元に戻すとする。このとき, 次の確率を求めなさい。

- (1) A さんが白玉を, B さんが白玉 1 個, 赤玉 1 個を取り出す確率。
 (2) B さんが白玉 1 個を取り出す確率。

例 1 とほぼ同じ問題になります。違いは取り出した玉を「元に戻す」という点。このことにより, **例 1** とどのような違いが出てくるのか確認していきましょう。

事象 A, B, C を次のように定める。

A : A さんが白玉を取り出す B : A さんが赤玉を取り出す

C : B さんが白玉 1 個, 赤玉 1 個を取り出す

- (1) 求める確率は

$$P(A \cap C) = P(A) \cdot P_A(C) = \frac{4}{6} \cdot \frac{2 \cdot 4}{{}_6C_2} = \frac{2}{3} \cdot \frac{8}{15} = \frac{16}{45}$$

- (2) 求める確率は

$$P(C) = P(A \cap C) + P(B \cap C) = \frac{16}{45} + \frac{2}{6} \cdot \frac{2 \cdot 4}{{}_6C_2} = \frac{8}{15}$$

本問の結果より, $P_A(C) = P(C)$ となっていることが分かる。これは,

「A という条件のもとでの C が起こる確率」= 「C が起こる確率」

ということなので, 「事象 C が起こる確率」が事象 A の影響を受けていないということである。

一般的に, 2 つの事象 A, B に対して, $P_A(B) = P(B)$ が成り立っているとき, 「事象 B は事象 A に独立である」といい, このとき以下の式が成り立つ。

$$P(A \cap B) = P(A) \cdot P(B)$$

なお, A さんが取り出す白玉の個数 X, B さんが取り出す白玉の個数を Y とすると, X, Y の同時確率分布は右のようになる。

ここで, 以下のように事象を定めると,

$$D : X = a \quad E : Y = b \quad (a = 0, 1, b = 0, 1, 2)$$

常に, $P(D \cap E) = P(D) \cdot P(E)$ が成り立っているので, 本問の事象 A, C に限らず,

「任意の a, b に対して, 事象 D, E が独立」

となっていることが分かる。

X \ Y	0	1	2	計
0	$\frac{1}{45}$	$\frac{8}{45}$	$\frac{2}{15}$	$\frac{1}{3}$
1	$\frac{2}{45}$	$\frac{16}{45}$	$\frac{4}{15}$	$\frac{2}{3}$
計	$\frac{1}{15}$	$\frac{8}{15}$	$\frac{2}{5}$	1

一般的に確率変数 X, Y に対し, X, Y のとる値をそれぞれ a, b とする。任意の a, b に対して,

$$P(X = a, Y = b) = P(X = a)P(Y = b)$$

が成り立っているとき, 確率変数 X と Y は**独立**であるという。

なお, 独立な 2 つの試行 S, T に対し, S の結果によって定まる確率変数 X と, T の結果によって定まる確率変数 Y は独立である。本問では, 取り出した玉を元に戻すことによって, A さんと B さんが玉を取り出す試行が独立になっているため, 確率変数 X, Y も独立になっている。

○独立な確率変数における積の期待値

2つの確率変数 X, Y に対して、積 XY も確率変数となる。ここでは X, Y が独立であるとき、確率変数 XY の期待値 $E(XY)$ について考える。

X, Y が独立なので、 $P(X = x_i) = p_i, P(Y = y_j) = q_j$ とすると、

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j) = p_i q_j$$

が成り立つ。これより、 $P(XY = x_i y_j) = p_i q_j$ となるので、

確率変数 X, Y の同時分布が右のようになる。

$X \backslash Y$	y_1	y_2	$\cdots$	y_n	計
x_1	$p_1 q_1$	$p_1 q_2$	$\cdots$	$p_1 q_n$	p_1
x_2	$p_2 q_1$	$p_2 q_2$	$\cdots$	$p_2 q_n$	p_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_m	$p_m q_1$	$p_m q_2$	$\cdots$	$p_m q_n$	p_m
計	q_1	q_2	$\cdots$	q_n	1

このとき、積 XY の期待値は以下のようになる。

$$\begin{aligned}
 E(XY) &= x_1 y_1 p_1 q_1 + x_1 y_2 p_1 q_2 + \cdots + x_1 y_n p_1 q_n \\
 &\quad + x_2 y_1 p_2 q_1 + x_2 y_2 p_2 q_2 + \cdots + x_2 y_n p_2 q_n \\
 &\quad + \cdots \\
 &\quad + x_m y_1 p_m q_1 + x_m y_2 p_m q_2 + \cdots + x_m y_n p_m q_n \\
 &= x_1 p_1 (y_1 q_1 + y_2 q_2 + \cdots + y_n q_n) \\
 &\quad + x_2 p_2 (y_1 q_1 + y_2 q_2 + \cdots + y_n q_n) + \cdots + x_m p_m (y_1 q_1 + y_2 q_2 + \cdots + y_n q_n) \\
 &= (x_1 p_1 + x_2 p_2 + \cdots + x_m p_m) (y_1 q_1 + y_2 q_2 + \cdots + y_n q_n) = E(X)E(Y)
 \end{aligned}$$

● 積の期待値 ●

独立な確率変数 X, Y に対して、以下の等式が成り立つ。

$$E(XY) = E(X)E(Y)$$

○独立な確率変数における和の分散

2つの確率変数 X, Y に対して、和 $X + Y$ も確率変数となる。ここでは X, Y が独立であるとき、確率変数 $X + Y$ の分散 $V(X + Y)$ について考える。

分散 $V(X)$ と期待値 $E(X)$ の関係式 $V(X) = E(X^2) - \{E(X)\}^2$ を用いると、

$$\begin{aligned}
 V(X + Y) &= E((X + Y)^2) - \{E(X + Y)\}^2 \\
 &= E(X^2 + 2XY + Y^2) - \{E(X) + E(Y)\}^2 \\
 &= E(X^2) + 2E(XY) + E(Y^2) - \{E(X)\}^2 - 2E(X)E(Y) - \{E(Y)\}^2 \\
 &= E(X^2) + 2E(X)E(Y) + E(Y^2) - \{E(X)\}^2 - 2E(X)E(Y) - \{E(Y)\}^2 \\
 &= E(X^2) - \{E(X)\}^2 + E(Y^2) - \{E(Y)\}^2 \\
 &= V(X) + V(Y)
 \end{aligned}$$

・以下の等式は常に成り立つ。
 $E(aX + bY) = aE(X) + bE(Y)$
 ・以下の等式は X, Y が独立であるとき成り立つ。
 $E(XY) = E(X)E(Y)$

● 和の分散 ●

独立な確率変数 X, Y に対して、以下の等式が成り立つ。

$$V(X + Y) = V(X) + V(Y)$$

なお、確率変数 X, Y, Z に対し、 X, Y, Z のとる値をそれぞれ a, b, c とする。任意の a, b, c に対して、

$$P(X = a, Y = b, Z = c) = P(X = a)P(Y = b)P(Z = c)$$

が成り立っているとき、確率変数 X, Y, Z は**独立**であるという。

このような独立である3つの確率変数についても上記と同様の関係式が成り立つ。

$$E(XYZ) = E(X)E(Y)E(Z)$$

$$V(X + Y + Z) = V(X) + V(Y) + V(Z)$$

§3 二項分布

1回の試行で事象 A が起こる確率が p とする。これを n 回行う反復試行において、A がちょうど k 回起こる確率は

$${}_nC_k p^k q^{n-k} \quad (\text{ただし, } q = 1 - p)$$

となる。このような試行において、事象 A が起こる回数を X とすると、 X は確率変数となり、その確率分布は以下のようになる。

X	0	1	2	...	n	計
P	${}_nC_0 p^0 q^n$	${}_nC_1 p q^{n-1}$	${}_nC_2 p^2 q^{n-2}$	...	${}_nC_n p^n q^0$	1

この表で与えられる確率分布を**二項分布 (binomial distribution)**といい、 $B(n, p)$ で表す。

○二項分布の期待値と分散

例5 さいころを3回投げるとき、1の目が出る回数の期待値と分散を求めなさい。

1の目が出る回数を X とすると、 X は二項分布 $B\left(3, \frac{1}{6}\right)$ に従います。

$P(X = k) = {}_3C_k \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{3-k}$ より、分布は右のようになります。

X	0	1	2	3	計
P	$\frac{125}{216}$	$\frac{25}{72}$	$\frac{5}{72}$	$\frac{1}{216}$	1

解①

$$E(X) = 0 \times \frac{125}{216} + 1 \times \frac{25}{72} + 2 \times \frac{5}{72} + 3 \times \frac{1}{216} = \frac{1}{2}$$

$$V(X) = E\left(\left(X - \frac{1}{2}\right)^2\right) = \left(0 - \frac{1}{2}\right)^2 \times \frac{125}{216} + \left(1 - \frac{1}{2}\right)^2 \times \frac{25}{72} + \left(2 - \frac{1}{2}\right)^2 \times \frac{5}{72} + \left(3 - \frac{1}{2}\right)^2 \times \frac{1}{216} = \frac{5}{12}$$

確率変数の積と和における期待値、分散の性質を使うと次のような解答になります。

解②

n 回目に1の目が出るとき1、1以外の目が出るとき0となるような確率変数を X_n ($n = 1, 2, 3$) とする。このとき、確率変数 X_n の期待値 $E(X_n)$ 、分散 $V(X_n)$ は以下のようになる。

$$E(X_n) = 0 \cdot \frac{5}{6} + 1 \cdot \frac{1}{6} = \frac{1}{6}$$

$$V(X_n) = E(X_n^2) - \{E(X_n)\}^2 = 0^2 \cdot \frac{5}{6} + 1^2 \cdot \frac{1}{6} - \left(\frac{1}{6}\right)^2 = \frac{1}{6} \left(1 - \frac{1}{6}\right) = \frac{1}{6} \cdot \frac{5}{6} = \frac{5}{36}$$

X_n	0	1	計
P	$\frac{5}{6}$	$\frac{1}{6}$	1

$X = X_1 + X_2 + X_3$ より、

$$E(X) = E(X_1 + X_2 + X_3) = E(X_1) + E(X_2) + E(X_3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

確率変数 X_1, X_2, X_3 は独立なので、

$$V(X) = V(X_1 + X_2 + X_3) = V(X_1) + V(X_2) + V(X_3) = \frac{5}{36} + \frac{5}{36} + \frac{5}{36} = \frac{5}{12}$$

一般的に、1回の試行で事象 A が起こる確率が p であるような試行を n 回繰り返したとき、事象 A が起こる回数の期待値と分散について考える。

k 回目に A が起こるとき 1, A が起こらないとき 0 となるような確率変数を X_k ($k = 1, 2, \dots, n$) とする。
このとき、確率変数 X_k の期待値 $E(X_k)$, 分散 $V(X_k)$ は以下ようになる。

$$E(X_k) = 0 \cdot q + 1 \cdot p = p$$

$$V(X_k) = E(X_k^2) - \{E(X_k)\}^2 = 0^2 \cdot q + 1^2 \cdot p - p^2 = p(1-p) = pq$$

X_k	0	1	計
P	q	p	1

(ただし, $q = 1 - p$)

$X = X_1 + X_2 + \dots + X_n$ より,

$$E(X) = E(X_1 + X_2 + \dots + X_n) = E(X_1) + E(X_2) + \dots + E(X_n) = p + p + \dots + p = np$$

確率変数 $X_1, X_2, \dots, X_n$ は独立なので,

$$V(X) = V(X_1 + X_2 + \dots + X_n) = V(X_1) + V(X_2) + \dots + V(X_n) = pq + pq + \dots + pq = npq$$

これより, $\sigma(X) = \sqrt{V(X)} = \sqrt{npq}$

二項分布の期待値・分散・標準偏差

確率変数 X の二項分布 $B(n, p)$ の期待値 $E(X)$, 分散 $V(X)$, 標準偏差 $\sigma(X)$ は以下ようになる。

$$E(X) = np, \quad V(X) = npq, \quad \sigma(X) = \sqrt{npq}$$

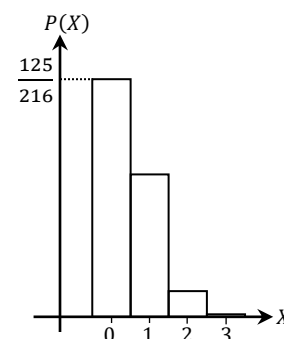
(ただし, $q = 1 - p$)

二項分布のヒストグラム

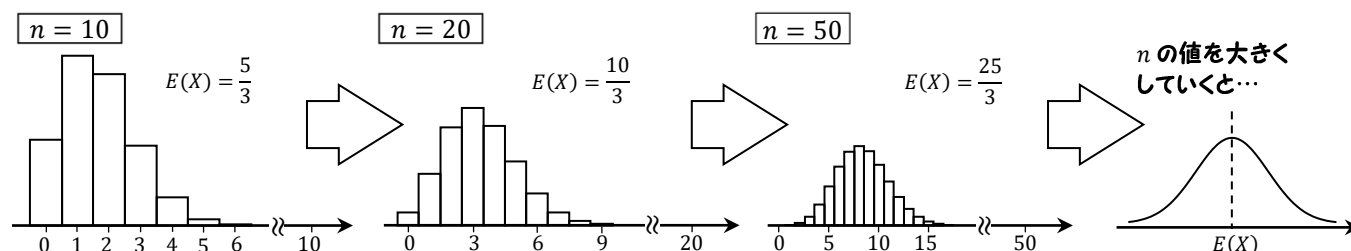
二項分布 $B\left(n, \frac{1}{6}\right)$ をヒストグラムで表すことを考える。

このとき, 例 5 におけるヒストグラム($n=3$ のとき)は右ようになる。

X	0	1	2	3	計
P	$\frac{125}{216}$	$\frac{25}{72}$	$\frac{5}{72}$	$\frac{1}{216}$	1



また, n の値を変化させたとき二項分布のヒストグラムは以下ようになる。



これを見ると, n の値を大きくしていくにつれて, ヒストグラムが期待値付近の値に関して左右対称の釣鐘型の曲線に近づいているように見える。



§ 4 正規分布

これまでの学んだ確率変数は離散的な値しかとらなかったが、実際は「長さ」や「重さ」など連続的な値も確率変数と考えることができる。ここでは、連続的な確率変数における確率分布について考えていく。

○連続型確率変数とその分布

右の表は、紙テープを目分量で 10cm の長さに切り、実際の長さを測るという実験を 3000 回行った結果である。

この 3000 本の紙テープの中から 1 本選ぶとき、その紙テープの長さを X とすると、 X が $8.75 \leq X < 10.25$ である確率は相対度数の 0.3027 であると考えてよい。このように考えると、 X が各階級に属する確率は相対度数に一致する。

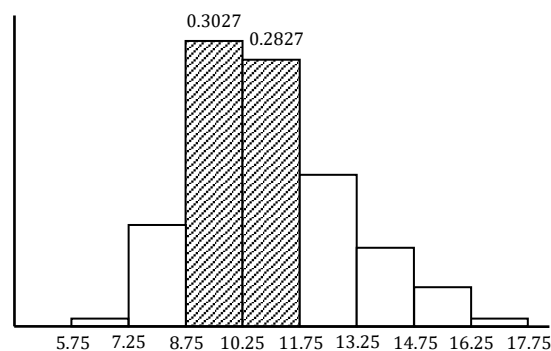
階級 (cm)	度数	相対度数
5.75以上 ~ 7.25未満	29	0.0097
7.25 ~ 8.75	326	0.1087
8.75 ~ 10.25	908	0.3027
10.25 ~ 11.75	848	0.2827
11.75 ~ 13.25	483	0.1610
13.25 ~ 14.75	250	0.0833
14.75 ~ 16.25	128	0.0427
16.25 ~ 17.75	28	0.0093
計	3000	1.0000

この表から、各階級の相対度数のヒストグラムをかくと右のようになる。このヒストグラムは

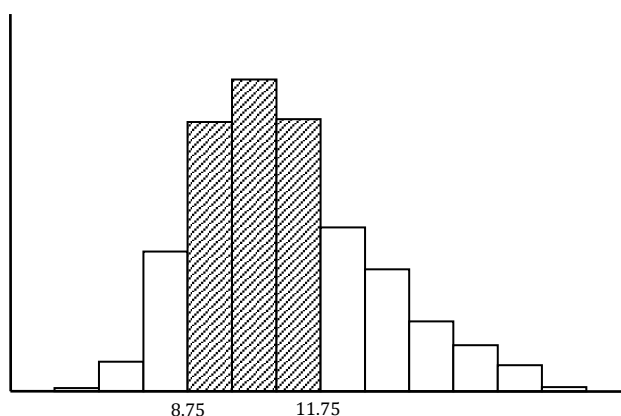
「各長方形の面積」=「相対度数」

となるように描かれている。つまり、 X が $8.75 \leq X < 11.75$ となる確率は右図の斜線部分の面積で表すことができる。なお、この長方形の面積の総和は 1 である。

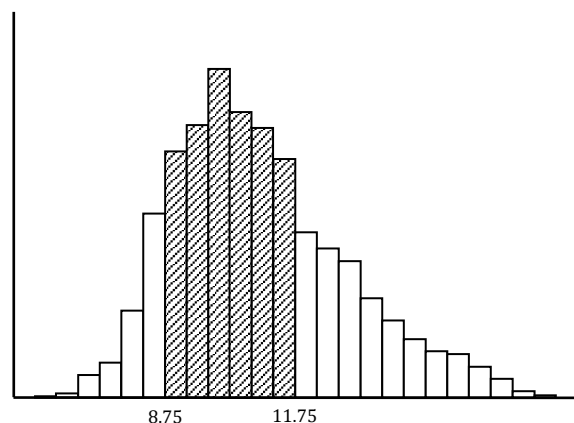
【図 1】のヒストグラムは階級の幅を 1.5 としているが、この幅を 1, 0.5 としたヒストグラムが以下の【図 2】【図 3】である。このとき、【図 1】と同様に、 X が $8.75 \leq X < 11.75$ となる確率は下図の斜線部分の面積となり、長方形の面積の総和は 1 である。



【図 1】階級の幅 1.5 のヒストグラム

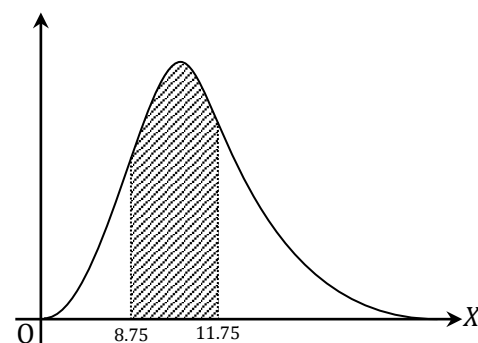


【図 2】階級の幅 1 のヒストグラム

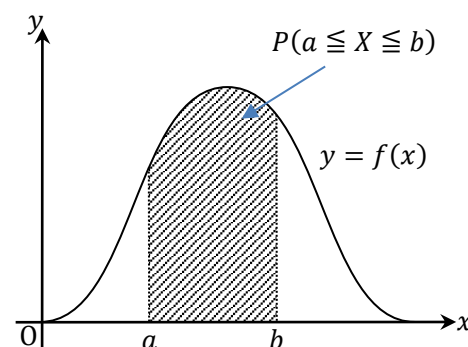


【図 3】階級の幅 0.5 のヒストグラム

この実験の回数をさらに増やしていき、階級の幅をさらに狭くしていくと、ヒストグラムの形は次第に右図のような曲線に近づいていくと考えられる。このとき、 X が $8.75 \leq X < 11.75$ となる確率は右図の斜線部分の面積となり、この曲線と X 軸で挟まれた部分の面積は 1 となる。



一般的に、連続的な値をとる確率変数を**連続型確率変数**という。連続型確率変数を考える場合は、確率変数 X に 1 つの曲線 $y = f(x)$ を対応させ、 $a \leq X \leq b$ となる確率 $P(a \leq X \leq b)$ が、右の図の斜線部分の面積で表されるようにする。この曲線 $y = f(x)$ を、 X の**分布曲線**といい、関数 $f(x)$ を**確率密度関数**という。このとき、 $y = f(x)$ と x 軸で挟まれた部分の面積は 1 となる。



確率密度関数 $f(x)$ には以下の性質がある。

確率密度関数の性質

- ① 常に $f(x) \geq 0$ である。
- ② $P(a \leq X \leq b) = \int_a^b f(x)dx$
- ③ X のとりうる値の範囲が $\alpha \leq X \leq \beta$ のとき $\int_{\alpha}^{\beta} f(x)dx = 1$

離散型確率変数 X において、 $P(X = x_k) = p_k$ ($k = 1, 2, \dots, n$) のとき、期待値 $E(X)$ 、分散 $V(X)$ は次のように表すことができた(ただし、 $E(x) = m$)。

$$E(X) = \sum_{k=1}^n x_k p_k, \quad V(X) = \sum_{k=1}^n (x_k - m)^2 p_k$$

連続型確率変数 X の場合、 X のとる値の範囲が $\alpha \leq X \leq \beta$ で、確率密度関数が $y = f(x)$ ($\alpha \leq x \leq \beta$) であるとき、 X の期待値 $E(X)$ 、分散 $V(X)$ は次のように表すことができる。

$$E(X) = \int_{\alpha}^{\beta} x f(x) dx, \quad V(X) = \int_{\alpha}^{\beta} (x - m)^2 f(x) dx$$

例6 確率変数 X のとりうる値の範囲が $0 \leq X \leq 3$ で、その確率密度関数が

$f(x) = \frac{2}{9}x(3-x)$ ($0 \leq x \leq 3$) で与えられているとき、次のものを求めなさい。

- (1) $1 \leq x \leq 2$ である確率 (2) 期待値 $E(X)$ (3) 分散 $V(X)$

$$(1) \quad P(1 \leq x \leq 2) = \int_1^2 \frac{2}{9}x(3-x)dx = \frac{2}{9} \left[\frac{3}{2}x^2 - \frac{1}{3}x^3 \right]_1^2 = \frac{2}{9} \left(\frac{9}{2} - \frac{7}{3} \right) = \frac{13}{27}$$

$$(2) \quad E(X) = \int_0^3 \frac{2}{9}x^2(3-x)dx = \frac{2}{9} \cdot \frac{1}{12}(3-0)^4 = \frac{3}{2}$$

$$(3) \quad V(X) = \int_0^3 \frac{2}{9}x \left(x - \frac{3}{2} \right)^2 (3-x)dx = \frac{2}{9} \int_0^3 x \left(x^2 - 3x + \frac{9}{4} \right) (3-x)dx$$

$$= \frac{2}{9} \int_0^3 \left\{ x^3(3-x) - 3x^2(3-x) + \frac{9}{4}x(3-x) \right\} dx = (*)$$

$$\int_0^3 x^3(3-x)dx = \int_0^3 (x-3+3)^3(3-x)dx$$

$$= \int_0^3 \{ -(x-3)^4 - 9(x-3)^3 - 27(x-3)^2 - 27(x-3) \} dx$$

$$= \left[-\frac{1}{5}(x-3)^5 - \frac{9}{4}(x-3)^4 - 9(x-3)^3 - \frac{27}{2}(x-3)^2 \right]_0^3 = \frac{243}{20}$$

$$\int_0^3 3x^2(3-x)dx = \frac{3}{12}(3-0)^4 = \frac{81}{4}$$

$$\int_0^3 \frac{9}{4}x(3-x)dx = \frac{9}{4} \cdot \frac{1}{6}(3-0)^3 = \frac{81}{8}$$

$$\text{以上より, } (*) = \frac{2}{9} \left(\frac{243}{20} - \frac{81}{4} + \frac{81}{8} \right) = \frac{9}{20}$$

1/12 公式

$$\int_{\alpha}^{\beta} a(x-\alpha)^2(x-\beta)dx = -\frac{a}{12}(\beta-\alpha)^4$$

1/6 公式

$$\int_{\alpha}^{\beta} a(x-\alpha)(x-\beta)dx = -\frac{a}{6}(\beta-\alpha)^3$$

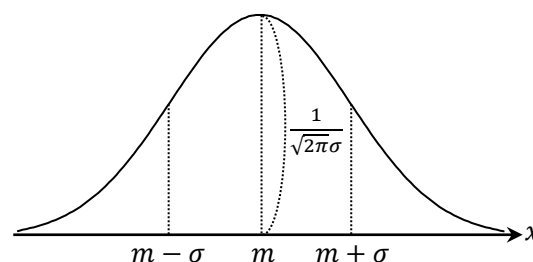
○正規分布

「目分量でテープを切る実験」のデータを連続型確率変数と考えたとき、その分布は**正規分布** (normal distribution) と呼ばれる分布になることが知られている。この他にも社会現象や自然現象のデータを連続型確率変数と考えたとき、その分布が正規分布に近くなるものが多数存在し、物事の起こる確率の計算をしたり、予測を立てたりするのに使うことができる。

連続型確率変数 X が正規分布に従うとき、確率密度関数 $f(x)$ は次の式で与えられ、分布曲線の概形は右図のようになる。

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

なお、式中の m は確率変数 X の期待値であり、 σ は標準偏差、 e は無理数の定数で、 $e = 2.71828 \dots$ である。



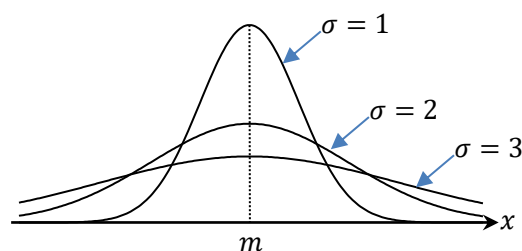
このとき、 X は**正規分布** $N(m, \sigma^2)$ に従うといい、曲線 $y = f(x)$ を**正規分布曲線**という。

● 正規分布の期待値・分散 ●

確率変数 X が正規分布 $N(m, \sigma^2)$ に従うとき、
期待値： $E(X) = m$ 、標準偏差： $\sigma(X) = \sigma$

また、正規分布曲線 $y = f(x)$ は、次のような性質を持つ。

- 【1】 直線 $x = m$ に関して対称で、 $x = m$ のとき最大値をとる。
- 【2】 標準偏差 σ が大きくなるほど平らな形になり、
 σ が小さくなるほど対称軸 $x = m$ のまわりに集まり、
山が高くなる。
- 【3】 x 軸を漸近線とする。



○標準正規分布

確率変数 X が正規分布 $N(m, \sigma^2)$ に従うとき、確率変数 $Z = \frac{X - m}{\sigma}$ も正規分布に従う。

$$Z \text{ の平均 : } E(Z) = E\left(\frac{X - m}{\sigma}\right) = \frac{1}{\sigma}E(X) - \frac{m}{\sigma} = \frac{m}{\sigma} - \frac{m}{\sigma} = 0$$

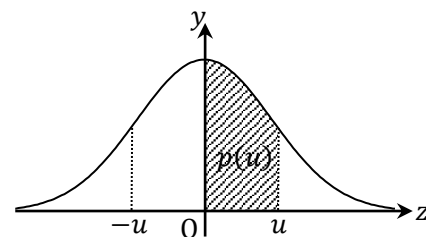
$$Z \text{ の分散 : } V(Z) = V\left(\frac{X - m}{\sigma}\right) = \frac{1}{\sigma^2}V(X) = \frac{1}{\sigma^2} \cdot \sigma^2 = 1$$

$$\begin{aligned} E(aX + b) &= aE(X) + b \\ V(aX + b) &= a^2V(X) \end{aligned}$$

より、確率変数 Z は正規分布 $N(0, 1)$ に従い、このとき Z の確率密度関数は $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ となる。

正規分布 $N(0, 1)$ を**標準正規分布**という。

標準正規分布 $N(0, 1)$ に従う確率変数 Z に対して、確率 $P(0 \leq z \leq u)$ を $p(u)$ とおく。このとき、様々な u の値に対する $p(u)$ の値を表にまとめたものが、p.16 の正規分布表である。



例えば、 $p(0.81)$ の値は、縦の列の「0.8」と横の列の「0.01」が交差したところに書いてある値 0.2910 になる。

また、確率密度関数 $f(z)$ は y 軸に関して対称であり、 $y = f(z)$ と x 軸で挟まれた部分の面積は 1 となるので、

$$P(z \geq 0) = P(z \leq 0), \quad P(z \geq 0) + P(z \leq 0) = 1$$

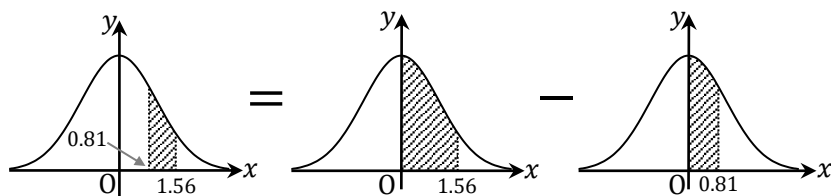
が成り立つ。これより、 $P(z \geq 0) = P(z \leq 0) = 0.5$ となる。

u	0.00	0.01	0.02
0.0	0.0000	0.0040	0.0080
0.1	0.0398	0.0438	0.0478
0.2	0.0793	0.0832	0.0871
0.3	0.1179	0.1217	0.1255
0.4	0.1554	0.1591	0.1628
0.5	0.1915	0.1950	0.1985
0.6	0.2257	0.2291	0.2324
0.7	0.2580	0.2611	0.2642
0.8	0.2881	0.2910	0.2939
0.9	0.3159	0.3186	0.3212
1.0	0.3413	0.3438	0.3461
1.1	0.3643	0.3665	0.3686
1.2	0.3849	0.3869	0.3888

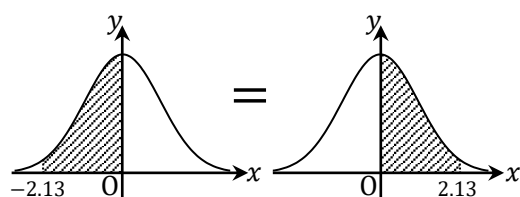
例7 標準正規分布に従う確率変数 X について、次の確率を求めなさい。

- (1) $P(0.81 \leq X \leq 1.56)$ (2) $P(-2.13 \leq X \leq 0)$ (3) $P(X \geq 1.87)$

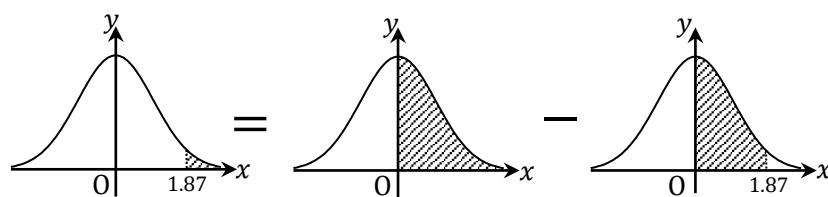
$$\begin{aligned} (1) \quad P(0.81 \leq X \leq 1.56) &= P(0 \leq X \leq 1.56) - P(0 \leq X \leq 0.81) \\ &= p(1.56) - p(0.81) = 0.4406 - 0.2910 = 0.1496 \end{aligned}$$



$$(2) \quad P(-2.13 \leq X \leq 0) = P(0 \leq X \leq 2.13) = p(2.13) = 0.4834$$



$$(3) \quad P(X \geq 1.87) = P(X \geq 0) - P(0 \leq X \leq 1.87) = 0.5 - p(1.87) = 0.5 - 0.4693 = 0.0307$$

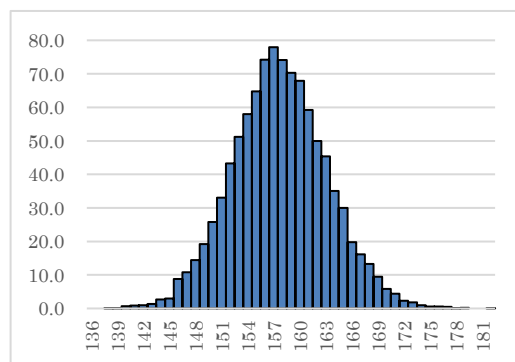


○正規分布の利用

「目分量で 10cm を測る」といったときに生じる測定誤差や、「人の身長」、「雨粒の大きさ」といった自然現象の中に現れるばらつきは正規分布になることが多いため、正規分布は統計学の中では最も重要かつ応用の幅が多い分布と言える。

右のグラフは 2021 年度の学校保健調査における 16 歳女子の身長の分布である。これを見てわかる通り、分布がほぼ正規分布と一致していることが分かる。

なお、2021 年度の 16 歳女子の身長の平均値は 157.7cm、標準偏差は 5.46cm である。これを利用して、次の確率を求めてみよう。



(単位は横軸 cm, 縦軸%)

例8 16 歳女子の中で身長が 160cm 以上の人は全体の約何%いますか。小数第 2 位を四捨五入して小数第 1 位まで求めなさい。

身長が正規分布に従うことを利用して解いていきます。

身長を X cm とすると、確率変数 X は正規分布 $N(157.7, 5.46^2)$ に従う。

このとき、 $Z = \frac{X - 157.7}{5.46}$ とおくと、確率変数 Z は正規分布 $N(0, 1)$ に従う。

$X = 160$ のとき、 $Z = \frac{160 - 157.7}{5.46} \approx 0.42$ なので

$$\begin{aligned} P(X \geq 160) &= P(Z \geq 0.42) = 0.5 - P(0 \leq Z \leq 0.42) \\ &= 0.5 - p(0.42) \\ &= 0.5 - 0.1628 = 0.3372 \end{aligned}$$

よって、約 33.7% いる。

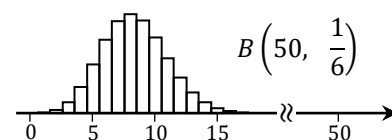
【正規分布表】

u	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

○二項分布の正規分布による近似

p.10 で見た通り、二項分布の n の値を大きくしていくと、そのヒストグラムは期待値付近の値に関して左右対称の釣鐘型の曲線に近づいていくが、実はこの曲線は正規分布曲線であることが知られている。

つまり、二項分布 $B(n, p)$ に従う確率変数 X は、 n が大きいとき、近似的に正規分布 $N(np, npq)$ に従うことが知られている。このことを利用すると、試行回数の多い反復試行の確率を近似的に求めることができる。



二項分布 $B(n, p)$ の期待値と標準偏差

$$E(X) = np, \quad \sigma(X) = \sqrt{npq} \quad (\text{ただし, } q = 1 - p)$$

◆ 二項分布と正規分布 ◆

n が大きいとき、二項分布 $B(n, p)$ に従う確率変数 X は、近似的に正規分布 $N(np, npq)$ に従う。

例9 1個のさいころを720回投げて、1の目が出る回数を X とする。このとき、 $100 \leq X \leq 150$ となる確率を近似的に求めなさい。

本来であれば、

$${}_{720}C_{100} \left(\frac{1}{6}\right)^{100} \left(\frac{5}{6}\right)^{620} + {}_{720}C_{101} \left(\frac{1}{6}\right)^{101} \left(\frac{5}{6}\right)^{619} + \cdots + {}_{720}C_{150} \left(\frac{1}{6}\right)^{150} \left(\frac{5}{6}\right)^{570}$$

を計算する問題ですが、ここでは正規分布を利用して近似的に求めます。

確率変数 X は二項分布 $B\left(720, \frac{1}{6}\right)$ に従う。

X の期待値 $E(X)$ と標準偏差 $\sigma(X)$ は

$$E(X) = 720 \cdot \frac{1}{6} = 120, \quad \sigma(X) = \sqrt{720 \cdot \frac{1}{6} \cdot \frac{5}{6}} = \sqrt{100} = 10$$

となる。試行回数が十分大きいので、確率変数 X は近似的に正規分布 $N(120, 100)$ に従うと考えてよい。

このとき、 $Z = \frac{X - 120}{10}$ は近似的に正規分布 $N(0, 1)$ に従うと考えてよい。

$X = 100$ のとき $Z = \frac{100 - 120}{10} = -2$, $X = 150$ のとき $Z = \frac{150 - 120}{10} = 3$ なので、

$$\begin{aligned} P(100 \leq X \leq 150) &= P(-2 \leq Z \leq 3) \\ &= P(-2 \leq Z \leq 0) + P(0 \leq Z \leq 3) \\ &= P(0 \leq Z \leq 2) + P(0 \leq Z \leq 3) \\ &= p(2) + p(3) = 0.4772 + 0.4987 = 0.9759 \end{aligned}$$

MEMO

統計的な推測

§1 母集団と標本

○全数調査と標本調査

統計的な調査には、調べたい対象全体のデータを集める**全数調査**と、調べたい対象の一部のデータを集め、全体を推測する**標本調査**の2通りの方法がある。

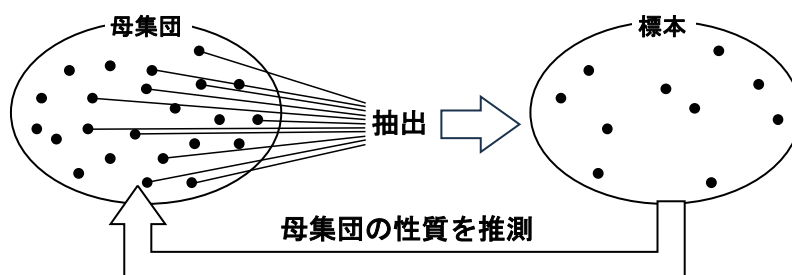
全数調査は、すべての対象を調べるので集まったデータに確実性があるというメリットがあるが、その分、データを集める労力、時間、費用がかかってしまうというデメリットがある。

それに対し標本調査は、一部のデータしか調べないので、労力、時間、費用がかからないというメリットがあるが、集めたデータに偏りがあつたり、用いた統計的な手法に誤りがあつたりすると、推測した結果が実際の結果と異なってしまうというデメリットがある。



例えば国政選挙では、最終的には全数調査を行い、当選・落選を決定するが、その前にマスコミによる出口調査と呼ばれる標本調査が行われ、開票結果が全て出そう前に当選・落選が報道される。出口調査は調査員が実際に投票所に行き調査をするが、一定の時間間隔や人数間隔で調査依頼をするなどの工夫をして偏りが出ないようにしている。このように、標本を抽出する際には何らかの手法で**無作為に抽出**できるように工夫する必要がある。

標本調査において、調べたい対象全体を**母集団**といい、調査のために母集団から抽出されたデータの集合を**標本**という。また、母集団に属する要素の総数を**母集団の大きさ**、標本に属する要素の総数を**標本の大きさ**という。また、無作為に抽出された標本を**無作為標本**という。



標本を抽出するとき、抽出したものを毎回元に戻しながら1個ずつ取り出すことを**復元抽出**といい、取り出したものをもとに戻さずに続けて抽出することを**非復元抽出**という。

母集団から n 個の標本を無作為に抽出し、その n 個の変量の値を $X_1, X_2, \dots, X_n$ とする。このとき $X_1, X_2, \dots, X_n$ は、「標本を抽出する」という試行によって定まる値なので確率変数となる。

復元抽出の場合、毎回、元の状態を復元してから取り出すので、確率変数 $X_1, X_2, \dots, X_n$ は互いに影響を受けないので独立となる。非復元抽出の場合は、元の状態に戻さないで確率変数 $X_1, X_2, \dots, X_n$ が独立になるとは限らない。

ただ、標本の大きさ n に比べて、母集団の大きさが十分に大きい場合は、非復元抽出であっても復元抽出と同様に確率変数が独立であるとしても構わない。

○母集団分布

大きさ N の母集団において、変量 x のとりうる値を

$$x_1, x_2, \dots, x_n$$

とし、それぞれの値をとる要素の個数(度数)を

$$f_1, f_2, \dots, f_n$$

とする。

x の値	度数
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_n	f_n
計	N

このとき、変量 x の値を X とすると、 X は確率変数となり、右のような確率分布に従う。このとき、各 X の値に対応する確率は変量 x の相対度数に一致する。

X	x_1	x_2	$\dots$	x_n	計
P	$\frac{f_1}{N}$	$\frac{f_2}{N}$	$\dots$	$\frac{f_n}{N}$	1

この X の確率分布を**母集団分布**という。また、確率変数 X の期待値と標準偏差を、それぞれ**母平均**、**母標準偏差**といい、 m 、 σ で表す。

平均値と期待値の違い

平均値と期待値はほとんど同じものですが、微妙に違います。

例えば、 n 個のデータ $x_1, x_2, x_3, \dots, x_n$ があつたとき、平均値とは

$$\frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

で求まる値です。それに対し、期待値とは n 個のデータ $x_1, x_2, x_3, \dots, x_n$ から 1 つを抽出しようとしているとき、その 1 つに期待できる値のことです。つまり、平均値とはすでに観測されたデータから求まる値であるのに対して、期待値はこれから観測するデータに対する理論値、予測値となります。

サイコロを振るという試行で考えると、1 回目に 2 の目が出て、2 回目に 4 の目が出た場合、その 2 回の試行の結果の平均値は

$$\frac{2 + 4}{2} = 3$$

となります。これを期待値とは言いません。

これに対し、期待値は出た目を X とすると、 X は確率変数で右のような確率分布に従うので、

$$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = \frac{7}{2}$$

X	1	2	3	4	5	6	計
P	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

となります。これはサイコロを振らなくても分かる理論値、予測値であり、サイコロを何回も振ったときの平均値が近づく値となります。

§ 2 標本平均の分布

母集団から n 個の標本を無作為に抽出し、その n 個の変量の値を $X_1, X_2, \dots, X_n$ とする。このとき

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

の値を**標本平均**といい、

$$S = \sqrt{\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X})^2}$$

を**標本標準偏差**という。

$X_1, X_2, \dots, X_n$ は、「標本を抽出する」という試行によって定まる値なので確率変数となる。

よって、 $\bar{X}, S$ も「標本を抽出する」という試行によって定まった $X_1, X_2, \dots, X_n$ から求められる値なので確率変数となる。

ここでは、標本平均 $\bar{X}$ の期待値 $E(\bar{X})$ 、分散 $S(\bar{X})$ 、標準偏差 $\sigma(\bar{X})$ について考えていく。

○1つの標本の期待値と分散

母平均 m 、母標準偏差 σ の母集団から 1 個の標本を抽出し、その変量の値を X_1 とする。

このとき X_1 の期待値 $E(X_1)$ 、標準偏差 $\sigma(X_1)$ は母平均、母標準偏差と一致する。

$$E(X_1) = m, \quad \sigma(X_1) = \sigma$$

これは母集団から抽出した 1 個の値の平均と標準偏差を求めるのではなく、1 個の値を抽出するとき、その 1 つに期待できる値と標準偏差を求めるということ。つまり、これから観測するデータに対する理論値、予測値を求めるということである。 X_1 という値は、母集団のすべての値をとりうるので、その期待値と標準偏差は母集団のものと一致する。

○標本平均の期待値と分散

母平均 m 、母標準偏差 σ の母集団から n 個の標本を無作為に抽出し、その n 個の変量の値を $X_1, X_2, \dots, X_n$ とする。このとき、各 X_k ($k = 1, 2, \dots, n$) の期待値と標準偏差は母平均、母標準偏差と一致する。

$$E(X_k) = m, \quad \sigma(X_k) = \sigma \quad (k = 1, 2, \dots, n)$$

これより、

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) \\ &= \frac{1}{n} \{E(X_1) + E(X_2) + \dots + E(X_n)\} \\ &= \frac{1}{n} (m + m + \dots + m) = \frac{1}{n} \cdot nm = m \end{aligned}$$

$$E(aX + bY) = aE(X) + bE(Y)$$

n 個の標本を復元抽出するとき、確率変数 $X_1, X_2, \dots, X_n$ は独立である。また、非復元抽出であっても n に比べ母集団の大きさが十分に大きいとき、確率変数 $X_1, X_2, \dots, X_n$ は独立であると考えてよい。よって、

$$\begin{aligned} V(\bar{X}) &= V\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) \\ &= \frac{1}{n^2} \{V(X_1) + V(X_2) + \dots + V(X_n)\} \end{aligned}$$

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

$$\sigma(\bar{X}) = \sqrt{V(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

- $V(aX + b) = a^2V(X)$
- X, Y が独立であるとき
 $V(X + Y) = V(X) + V(Y)$

🎯 標本平均の期待値・分散・標準偏差 🎯

母平均 m 、母標準偏差 σ の母集団から n 個の標本を無作為に抽出するとき、標本平均 $\bar{X}$ の期待値 $E(\bar{X})$ 、分散 $S(\bar{X})$ 、標準偏差 $\sigma(\bar{X})$ は、

$$E(\bar{X}) = m, \quad V(\bar{X}) = \frac{\sigma^2}{n}, \quad \sigma(\bar{X}) = \sqrt{V(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

この結果から分かる通り、 n の値が大きくなればなるほど標準偏差 $\sigma(\bar{X})$ の値は小さくなる。つまり、抽出する標本の大きさが大きいほど、真の平均(母平均)に近い値が求まるということである。このことは、以下のような具体例で考えるとよい。

例えば、300 人が受けたテストの平均を調べるために、10 人を抽出し平均を求めると、その値は 10 人の選び方によって、真の平均(母平均)に近い値が出てくる可能性もあるし、全く違う値が出てくる可能性もある。ただ、300 人の中から 250 人を抽出し平均を求めると、それは真の平均にかなり近い値になるはずである。

なお、これは「**大数の法則**」と呼ばれるとても重要な法則である。



🎯 大数の法則 🎯

母集団から大きさ n の標本を無作為に抽出するとき、その標本平均 $\bar{X}$ は、 n が大きくなるに従って、母平均に近づく。

○標本平均の分布

標本平均 $\bar{X}$ の分布について次のことが知られている。

◆◆ 中心極限定理 ◆◆

母平均 m ，母標準偏差 σ の母集団から大きさ n の標本を無作為に抽出するとき，標本平均 $\bar{X}$ は， n が大きいとき近似的に正規分布 $N\left(m, \frac{\sigma^2}{n}\right)$ に従う。

この定理は，母集団の分布にかかわらず，標本の大きさが十分に大きい場合，標本平均 $\bar{X}$ の分布は正規分布に近づいていくという定理であるが，母集団分布が正規分布であるときは， n が大きくなくても，常に $\bar{X}$ は正規分布に従う。

例10 母平均 50，母標準偏差 20 の母集団から，大きさ 100 の標本を無作為に抽出する。
このとき，標本平均 $\bar{X}$ が 45 より小さい値をとる確率を求めなさい。

母平均 m ，標本平均 $\bar{X}$ ，母標準偏差 σ とおくと，標本の大きさ $n = 100$ は十分に大きいので， $\bar{X}$ は正規分布 $N\left(m, \frac{\sigma^2}{n}\right)$ に従うと考えて構いません。

標本平均 $\bar{X}$ は近似的に正規分布 $N\left(50, \frac{20^2}{100}\right)$ ，つまり $N(50, 4)$ に従う。

ここで， $Z = \frac{\bar{X} - 50}{2}$ とおくと， Z は近似的に標準正規分布 $N(0, 1)$ に従う。

$\bar{X} = 45$ とすると， $Z = -2.5$

これより，

$$P(\bar{X} < 45) = P(Z < -2.5) = P(Z > 2.5) = 0.5 - p(2.5) = 0.5 - 0.4938 = 0.0062$$

この結果から分かる通り，大きさ 100 の標本を抽出すると，標本平均 $\bar{X}$ が 45 より小さい値をとる確率がわずか 6%しかありません。このことから，「標本の大きさが大きくなればなるほど母平均に近づく」という「大数の法則」の正しさが分かるかと思います。

○標本比率と正規分布

母集団全体の中で、ある特性 A をもつ要素の割合を、特性 A の**母比率**といい、この母集団の中から抽出した標本の中で、特性 A をもつ要素の割合を、特性 A の**標本比率**という。

例えば、

「ある高校では身長が 170cm 以上の生徒が全体の 25%いるとする。この集団の中から 100 人の生徒を無作為に抽出し、身長を調べたところ 170cm 以上の生徒が 20 人いた。」

という場合、母比率は 0.25、標本比率は 0.2 となる。

特性 A の母比率が p である母集団から、大きさ n の標本を無作為に抽出する。

この中で特性 A を持つものの個数を T とすると、 T は確率変数となる。

n 個の標本の中で特性 A を持つものの数が k となる確率は

$$P(T = k) = {}_nC_k p^k q^{n-k} \quad (\text{ただし, } q = 1 - p)$$

となるので、 T は二項分布 $B(n, p)$ に従う。

T	0	1	2	...	n	計
P	${}_nC_0 p^0 q^n$	${}_nC_1 p^1 q^{n-1}$	${}_nC_2 p^2 q^{n-2}$	...	${}_nC_n p^n q^0$	1

n が大きいとき T は近似的に正規分布 $N(np, npq)$ に従う。

このとき、標本比率を R とすると、 $R = \frac{T}{n}$ となるので、

$$E(R) = E\left(\frac{T}{n}\right) = \frac{1}{n} E(T) = \frac{1}{n} \cdot np = p \quad V(R) = V\left(\frac{T}{n}\right) = \frac{1}{n^2} \cdot npq = \frac{pq}{n}$$

$$\begin{aligned} E(aX + b) &= aE(X) + b \\ V(aX + b) &= a^2 V(X) \end{aligned}$$

これより、標本比率 R は近似的に正規分布 $N\left(p, \frac{pq}{n}\right)$ に従うことが分かる。

🔗 標本比率と正規分布 🔗

特性 A の母比率が p である母集団から、大きさ n の標本を無作為に抽出するとき、標本比率 R は、 n が大きいとき近似的に正規分布 $N\left(p, \frac{pq}{n}\right)$ に従う。

なお、以下のように考えると、標本比率が正規分布に従うことは中心極限定理を用いて示すこともできる。

大きさ n の無作為標本 1 つ 1 つに対して、 $X_1, X_2, \dots, X_n$ を次のように定める。

特性 A をもつとき $X_k = 1 \quad (k = 0, 1, \dots, n)$

特性 A をもたないとき $X_k = 0 \quad (k = 0, 1, \dots, n)$

このとき、 $T = X_1 + X_2 + \dots + X_n$ となるので、標本比率 R は

$$R = \frac{T}{n} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

と表すことができる。これは、標本比率 R は標本平均 $\bar{X}$ とみなすことができる。

中心極限定理より、標本平均 $\bar{X}$ は n が大きいとき近似的に正規分布に従うので、標本比率 R も n が大きいとき近似的に正規分布に従うことが分かる。

例11 1個のさいころを320回投げるとき、1の目が出る相対度数を R とする。このとき、

$$P\left(\left|R - \frac{1}{6}\right| \leq 0.04\right) \text{ を求めなさい。}$$

相対度数 R は $\frac{(1 \text{ の目が出た回数})}{320}$ なので標本比率です。ここでは、 R が正規分布に従うことを利用します。

相対度数 R は近似的に正規分布 $N\left(\frac{1}{6}, \frac{\frac{1}{6} \cdot \frac{5}{6}}{320}\right)$, つまり $N\left(\frac{1}{6}, \frac{1}{48^2}\right)$ に従う。

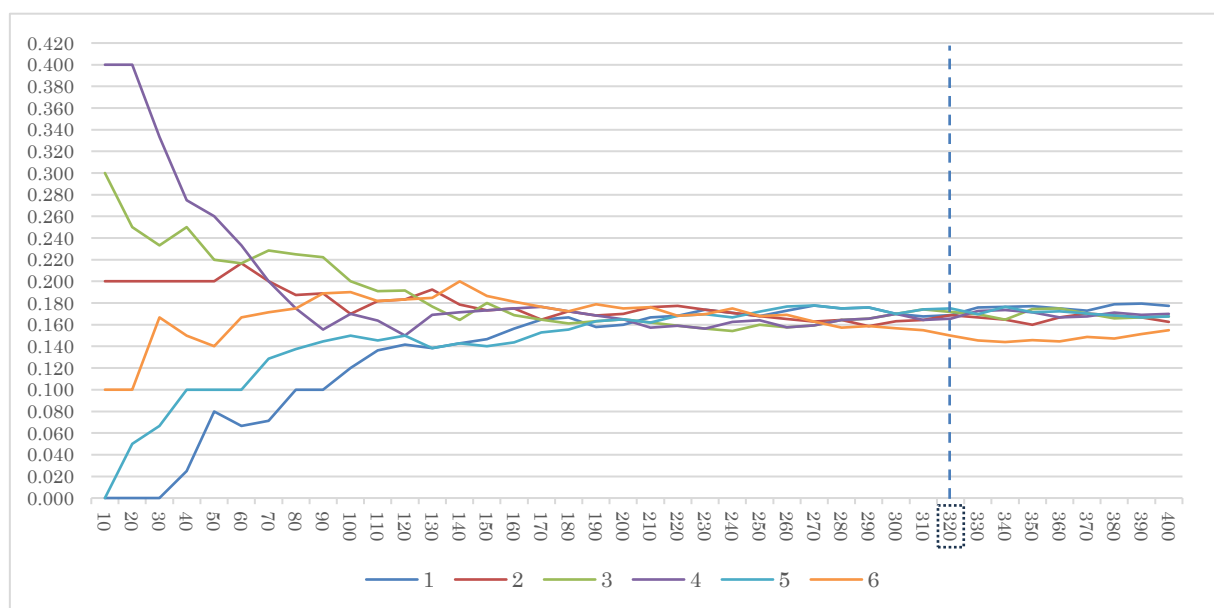
ここで、 $Z = \frac{R - \frac{1}{6}}{\frac{1}{48}} = 48R - 8$ とおくと、 Z は近似的に標準正規分布 $N(0, 1)$ に従う。

$$\left|R - \frac{1}{6}\right| \leq 0.04 \Leftrightarrow -0.04 \leq R - \frac{1}{6} \leq 0.04 \Leftrightarrow \frac{1}{6} - 0.04 \leq R \leq \frac{1}{6} + 0.04$$

$R = \frac{1}{6} + 0.04$ のとき $Z = 1.92$, $R = \frac{1}{6} - 0.04$ のとき $Z = -1.92$ となるので、

$$\begin{aligned} P\left(\left|R - \frac{1}{6}\right| \leq 0.04\right) &= P(-1.92 \leq Z \leq 1.92) = 2P(0 \leq Z \leq 1.92) \\ &= 2p(1.92) = 2 \times 0.4726 = 0.9452 \end{aligned}$$

この結果を見て分かる通り、320回サイコロを振ると、「1の目が出る相対度数」と「1の目が出る確率」の差がかなりの高確率で0.04以下になるということです。実際に、サイコロを400回振り、各目の出る確率の変化を10回おきに計測してみると以下のような結果になります。回数を増やしていくと、各目の出る確率が急激に $\frac{1}{6} = 0.16666 \dots$ に近づいていることが分かります。



§3 推定

ここでは、取り出した標本から母集団の性質を推定する方法を学んでいく。

○母平均の推定

例12 ある地域の16歳100人を無作為に抽出し、平均身長を調べたところ168cmであった。
母標準偏差を5とすると、その地域の16歳の平均身長はどのくらいですか。



「大数の法則」によると、標本の大きさが大きくなればなるほど、
標本平均は母平均に近づいていくので、100人も(?) 抽出したのであれば、
標本平均168cmは母平均にかなり近いことが予想されます。
では、いったいどれほど近いのか？

【大数の法則】
母集団から大きさ n の標本を
無作為に抽出するとき、その
標本平均 $\bar{X}$ は、 n が大きくな
るに従って、母平均に近づく。

母平均 m ，母標準偏差 σ ，標本平均 $\bar{X}$ とする。

このとき、 $\bar{X}$ は正規分布 $N\left(m, \frac{\sigma^2}{n}\right)$ に従う。

$Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}}$ とおくと、 Z は標準正規分布 $N(0, 1)$ に従う。

正規分布表を用いると、

$$\begin{aligned} P(|Z| \leq 1.96) &= P(-1.96 \leq Z \leq 1.96) \\ &= 2P(0 \leq Z \leq 1.96) = 2p(1.96) = 2 \times 0.4750 = 0.95 \quad \cdots \textcircled{1} \end{aligned}$$

$$\begin{aligned} \text{また、} P(|Z| \leq 1.96) &= P\left(\left|\frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}}\right| \leq 1.96\right) \\ &= P\left(-1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X} - m \leq 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \\ &= P\left(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) \quad \cdots \textcircled{2} \end{aligned}$$

$$\textcircled{1}, \textcircled{2} \text{より、} P\left(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

これより、母平均 m は95%の確率で区間 $\left[\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right]$ に含まれる。

$\bar{X} = 168$, $\sigma = 5$, $n = 100$ より、母平均 m は95%の確率で

区間 $\left[168 - 1.96 \cdot \frac{5}{10}, 168 + 1.96 \cdot \frac{5}{10}\right]$, つまり $[167.02, 168.98]$ に含まれる。

【中心極限定理】

母平均 m ，母標準偏差 σ の母集団から
大きさ n の標本を無作為に抽出する
とき、標本平均 $\bar{X}$ は、 n が大きいとき近
似的に正規分布 $N\left(m, \frac{\sigma^2}{n}\right)$ に従う。

母平均 m ，母標準偏差 σ の母集団から，大きさ n の標本を無作為抽出したとき，標本平均を $\bar{X}$ とする。
このとき一般的に，以下の式が成り立つ。

$$P\left(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq m \leq \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

つまり，母平均 m は 95% の確率で区間 $\left[\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right]$ に含まれるということである。

この区間を母平均 m に対する**信頼度 95%の信頼区間**という。

なお，推定の精度をもう少し高めたい場合は信頼度 99%で推定することがある。
確率分布 Z が標準正規分布 $N(0, 1)$ に従うとき，

$$P(|Z| \leq 2.58) = P(-2.58 \leq Z \leq 2.58) = 2P(0 \leq Z \leq 2.58) = 2p(2.58) = 2 \times 0.4951 \approx 0.99$$

となるので，**例 12**と同様に考えると，母平均 m の**信頼度 99%の信頼区間**は以下のようになる。

$$\left[\bar{X} - 2.58 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + 2.58 \cdot \frac{\sigma}{\sqrt{n}}\right]$$

これを用いて，**例 12**の母平均の信頼度 99%の信頼区間を求めると，

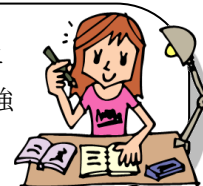
$$\left[168 - 2.58 \cdot \frac{5}{10}, 168 + 2.58 \cdot \frac{5}{10}\right] \text{ つまり } [166.71, 169.29]$$

となる。95%の信頼区間に比べると少し広がっていることが分かる。

母標準偏差が分からない場合

例 12では母標準偏差を用いて推定しましたが，母平均が分からないのに，母標準偏差は分かっているというケースはほとんどありません。ただ，**標本の大きさ n が十分に大きければ，母標準偏差の代わりに標本標準偏差を用いて母平均を推定することができます。**

例13 ある地域の 16 歳 1600 人を無作為に抽出し，平日の学校以外の勉強時間を調べたところ平均値 80 分，標準偏差が 60 分であった。このとき，全国の 16 歳の平均勉強時間を信頼度 95%で推定しなさい。



母平均 m ，標本平均 $\bar{X}$ ，標本標準偏差 S とおくと，標本の大きさ $n = 1600$ は十分に大きいので，

$\bar{X}$ は正規分布 $N\left(m, \frac{S^2}{n}\right)$ に従うと考えて構いません。

標本平均 $\bar{X}$ ，標本標準偏差 S とおくと，標本の大きさ n は十分に大きいので，

母平均 m は 95% の確率で区間 $\left[\bar{X} - 1.96 \cdot \frac{S}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{S}{\sqrt{n}}\right]$ に含まれる。

これより， $\bar{X} = 80$ ， $S = 60$ ， $n = 1600$ より，母平均 m は 95% の確率で

区間 $\left[80 - 1.96 \cdot \frac{60}{40}, 80 + 1.96 \cdot \frac{60}{40}\right]$ ，つまり $[77.06, 82.94]$ に含まれる。

○母比率の推定

ここでは、標本比率から母比率を推定する方法を考える。

以前学んだ通り、母比率を p とすると標本比率 R は近似的に正規分布 $N\left(p, \frac{p(1-p)}{n}\right)$ に従う。

ここで、 $Z = \frac{R-p}{\sqrt{\frac{p(1-p)}{n}}}$ とおくと、 Z は近似的に標準正規分布 $N(0, 1)$ に従うので、

$$\begin{aligned} P(|Z| \leq 1.96) = 0.95 &\Leftrightarrow P\left(\left|\frac{R-p}{\sqrt{\frac{p(1-p)}{n}}}\right| \leq 1.96\right) = 0.95 \\ &\Leftrightarrow P\left(R - 1.96 \cdot \sqrt{\frac{p(1-p)}{n}} \leq p \leq R + 1.96 \cdot \sqrt{\frac{p(1-p)}{n}}\right) = 0.95 \end{aligned}$$

標本比率 R は標本平均 $\bar{X}$ の特別な場合と考えられるので(p.24)、大数の法則より、 n が十分に大きいとき母比率 p は標本比率 R に近い値になる。

これより、 p を R に置き換えると、

$$P\left(R - 1.96 \cdot \sqrt{\frac{R(1-R)}{n}} \leq p \leq R + 1.96 \cdot \sqrt{\frac{R(1-R)}{n}}\right) = 0.95$$

つまり、母比率 p に対する信頼度 95% の信頼区間は

$$\left[R - 1.96 \cdot \sqrt{\frac{R(1-R)}{n}}, R + 1.96 \cdot \sqrt{\frac{R(1-R)}{n}} \right]$$

となる。

【大数の法則】

母集団から大きさ n の標本を無作為に抽出するとき、その標本平均 $\bar{X}$ は、 n が大きくなるに従って、母平均に近づく。

例14 ある学校の生徒 400 人を無作為に調査したところ、225 人が東京都から通学していた。

このとき、この学校全体の東京都から通っている生徒の母比率 p を信頼度 95% で推定しなさい。



標本比率は $R = \frac{225}{400} = \frac{9}{16}$ となる。

$$\text{このとき、} 1.96 \cdot \sqrt{\frac{R(1-R)}{n}} = 1.96 \cdot \sqrt{\frac{\frac{9}{16}(1-\frac{9}{16})}{400}} = 1.96 \cdot \frac{3\sqrt{7}}{320} \approx 0.049$$

これより、母比率 p は

区間 $\left[\frac{9}{16} - 0.049, \frac{9}{16} + 0.049\right]$ 、つまり $[0.5135, 0.6115]$ に含まれる。

§4 仮説検定

ある食堂にはAランチとBランチの2種類のランチセットがある。ある日、「AランチとBランチの人気に差はあるか？」を調べるために、400人にアンケートをとったところ、Aランチと答えた人が221人、Bランチと答えた人が179人であった。この結果から

AランチとBランチの人気に差がある

と判断してよいだろうか？



ここではまず、「AランチとBランチの人気に差はない」と仮定して考えていく。「差はない」と仮定すると、「Aランチ」と答える確率と、「Bランチ」と答える確率は、どちらも0.5となる。このとき、Aランチと答える人が221人になる確率は、

$${}_{400}C_{221}(0.5)^{221}(0.5)^{179}$$

となるが、この値を計算するのではなく、ここでは「Aランチ」と答える人数 X が二項分布 $B(400, 0.5)$ に従う確率変数であることを利用する。

X は二項分布に従うので、期待値 $E(X)$ と標準偏差 $\sigma(X)$ は

$$E(X) = 400 \cdot 0.5 = 200, \sigma(X) = \sqrt{400 \cdot 0.5 \cdot 0.5} = \sqrt{100} = 10$$

となる。

また、試行回数が十分大きいので、確率変数 X は近似的に正規分布 $N(200, 100)$ に従うと考えてよい。

二項分布 $B(n, p)$ の期待値と標準偏差

$$E(X) = np, \sigma(X) = \sqrt{npq} \quad (\text{ただし, } q = 1 - p)$$

二項分布 $B(n, p)$ に従う確率変数 X は、 n が大きいとき、近似的に正規分布 $N(np, npq)$ に従う。

このとき、 $Z = \frac{X - 200}{10}$ は正規分布 $N(0, 1)$ に従う。

$X = 221$ のとき $Z = \frac{221 - 200}{10} = 2.1$ なので、 $X \geq 221$ となる確率は、

$$P(X \geq 221) = P(Z \geq 2.1) = 0.5 - P(0 \leq Z \leq 2.1) = 0.5 - p(2.1) = 0.5 - 0.4821 = 0.0179$$

となる。つまり、「AランチとBランチの人気に差はない」という仮定のもとで、「Aランチ」と回答する人が221人以上になる可能性は1.79%ととても低いことが分かる。このことから考えられる結論としては、

【結論1】「AランチとBランチの人気に差はない」という仮定は正しく、かつ可能性が低い(1.79%)ことが偶然に起こった。

【結論2】「AランチとBランチの人気に差はない」という仮定がそもそも正しくない。

の2つになるが、この場合は[結論2]、つまり

AランチとBランチの人気に差がある(Aランチの方がBランチより人気がある)

を採用した方が適切と思われる。

一般的に、ある母集団に関する仮定を**仮説**といい、この仮説が正しいかどうかを確率で判断する方法を**仮説検定**という。また、上の例のように、仮説が正しくないと判断して、その仮説を採用しないことを、仮説を**棄却する**という。

○仮説を立てる際のポイント

仮説検定では必ず「〇〇と△△の差はない」というような仮説を立てるところから始める。「〇〇と△△には差がある」とすると、「大きな差」や「わずかな差」など「差」の度合いがいろいろと考えられてしまうが、「差はない」とすれば、その状況は1通りしかなく、先ほどの例のように「確率 0.5」として計算することが可能となるからである。

「〇〇と△△の差はない」というような仮説を**帰無仮説**といい、 H_0 と表す。それに対し、本来正しいかどうか判断したい「〇〇と△△には差がある」というような仮説を**対立仮説**といい、 H_1 と表す。

仮説検定を行う際にはまず、帰無仮説と対立仮説の設定するところから始める。

○帰無仮説を棄却する際のポイント

先ほどの例では「帰無仮説のもとでその事象が起こる確率」が 1.79%と低かったので帰無仮説を棄却したが、一般的には、「帰無仮説のもとでその事象が起こる確率」が 5%より低い場合に「稀に起こること」と判断して帰無仮説を棄却するのが慣例である。なお、さらに厳密な判定が必要な場合には、1%やもっと低い確率を利用することもある。

仮説検定では、5%、1%のように、あらかじめ基準となる確率を決めておき、それより「帰無仮説のもとでその事象が起こる確率」が小さい場合に棄却する。この基準となる確率を**有意水準**という。

また、有意水準 α を決めると、それに対して仮説が棄却されるような確率変数の値の範囲も定まる。この範囲のことを有意水準 α の**棄却域**といい、仮説検定を行う際には標本から得られる確率変数の値が、この棄却域に含まれるかどうかで判断する。

仮説検定の手順をまとめると次のようになる。

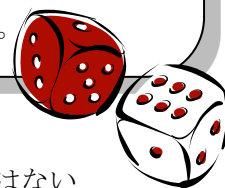
❖ 仮説検定の手順 ❖

1. 帰無仮説 H_0 と対立仮説 H_1 を設定する。
2. 帰無仮説 H_0 が正しいと仮定し、対象となる統計量の分布を調べる。
3. 有意水準 α を定め、棄却域を求める。
4. 標本から得られた確率変数の値が棄却域に含まれるかを調べ、棄却域に含まれていれば H_0 を棄却する。

仮説検定では、本来正しいかどうか判断したい対立仮説に設定する。これは証明したい命題の否定を仮定し、矛盾を導くことで証明する背理法の手法と同じである。

○母比率の検定

例15 引き出しの奥にあった古いサイコロを 180 回投げたところ 1 の目が 40 回出ました。
このサイコロは古いのでゆがみがあるのではないかと。有意水準 5% で検定しなさい。



この場合

対立仮説 H_1 : サイコロにゆがみがある 帰無仮説 H_0 : サイコロにゆがみはない
となります。「サイコロにゆがみがない」とすると 1 の目が出る確率が $\frac{1}{6}$ となり、そのもとで、
1 の目が 40 回出る確率について検証していきます。

1 の目が出る確率を p とすると、

サイコロにゆがみがあるならば $p \neq \frac{1}{6}$ となる。

ここで、サイコロにゆがみがない、つまり $p = \frac{1}{6}$ であるという仮説を立てる。

この仮説の下で、サイコロを 180 回投げたとき、1 の目が出る回数 X は二項分布 $B\left(180, \frac{1}{6}\right)$ に従う。

よって、 X の期待値、標準偏差は、 $E(X) = 180 \cdot \frac{1}{6} = 30$, $\sigma(X) = \sqrt{180 \cdot \frac{1}{6} \cdot \frac{5}{6}} = 5$

これより、 $Z = \frac{X - 30}{5}$ は近似的に標準正規分布 $N(0, 1)$ に従う。

← 標本の大きさが十分に大きいとき、
 X は $N(30, 5^2)$ に従うので。

正規分布表より $P(-1.96 \leq Z \leq 1.96) \approx 0.95$ なので、
有意水準 5% の棄却域は $Z \leq -1.96$, $Z \geq 1.96$ となる。

$X = 40$ のとき、 $Z = \frac{40 - 30}{5} = 2$ であり、この値は棄却域に入るから仮説を棄却できる。

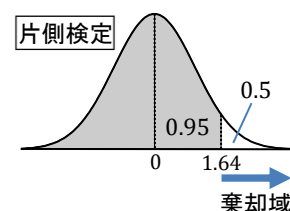
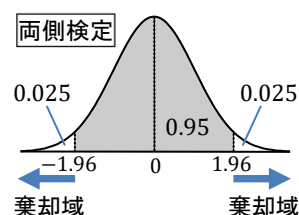
したがって、 $p \neq \frac{1}{6}$ つまりサイコロにゆがみがあると判断してよい。

- もし「有意水準 1% で検定しなさい」だったら、正規分布表より $P(-2.58 \leq Z \leq 2.58) \approx 0.99$ なので、
棄却域は $Z \leq -2.58$, $Z \geq 2.58$ となる。このとき、 $Z = 2$ は棄却域に入らないので、棄却することは
できない。一般的に有意水準が 1% より 5% の方が棄却しやすいが、棄却しやすいからといって安易に
5% を選ぶものではない。その検定する対象に適した有意水準を設定しなくてはならない。

- この例では、「サイコロにゆがみがあるか？」を知りたいので、
1 の目が出る回数が多すぎても少なすぎても仮説が棄却できるように、
棄却域を両側にとって考えている。このような検定を**両側検定**という。

これに対し、「1 の目が出やすいか？」を知りたい場合は、
1 の目が出る回数が多すぎるときに仮説が棄却されるように、
棄却域を片側だけにとる。このような検定を**片側検定**という。

【有意水準 5% の場合】



例16 Aさんは体育の授業で鉄棒を行ったところ、3回に1回の割合で「逆上がり」が成功する。その後、猛練習に励んだところ、162回中63回成功した。この結果は練習の効果と判断してよいだろうか。有意水準5%で検定しなさい。



この場合

対立仮説 H_1 : 練習の効果があった 帰無仮説 H_0 : 練習の効果がなかった

となります。ここでは、「練習の効果」について検証したいので、成功回数が多すぎるときに仮説が棄却されるように片側検定を行います。

「逆上がり」が成功する確率を p とすると、

練習に効果があったならば $p > \frac{1}{3}$ となる。

ここで、練習に効果がなかった、つまり $p = \frac{1}{3}$ であるという仮説を立てる。

この仮説の下で、162回「逆上がり」を行ったとき、成功する回数 X は二項分布 $B\left(162, \frac{1}{3}\right)$ に従う。

このとき、 X の期待値、標準偏差は、 $E(X) = 162 \cdot \frac{1}{3} = 54$, $\sigma(X) = \sqrt{162 \cdot \frac{1}{3} \cdot \frac{2}{3}} = 6$

これより、 $Z = \frac{X - 54}{6}$ は近似的に標準正規分布 $N(0, 1)$ に従う。

←標本の大きさが十分に大きいとき、
 X は $N(54, 6^2)$ に従うので。

正規分布表より $P(Z \leq 1.64) \approx 0.5 + 0.45 = 0.95$ なので、

有意水準5%の棄却域は $Z \geq 1.64$ となる。

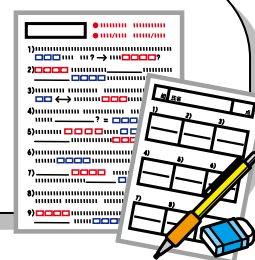
$X = 63$ のとき、 $Z = \frac{63 - 54}{6} = \frac{9}{6} = 1.5$ であり、この値は棄却域に入らないから仮説は棄却できない。

したがって、この結果からは練習に効果があったとは判断できない。

- ・棄却できないからと言って、帰無仮説「練習の効果がなかった」が採択できるわけではない。この検定では帰無仮説を棄却する根拠が標本から得られなかったというだけである。
- ・仮説検定を通して何か結論が得られるのは、確率変数の値が棄却されるときに限られる。つまり、棄却域の範囲次第で、確率変数の値に意味があるかないかが決まるので、棄却域の範囲を定める水準を有意水準と呼ぶ。
- ・例えば有意水準が5%のとき、5%の確率で正しい仮説を誤って棄却してしまう危険がある。そのため、有意水準のことを危険率と呼ぶこともある。

○母平均の検定

例17 ある模試を受けた T 高校の生徒の中から無作為に 100 人を抽出したところ、平均点 57 点、標準偏差 15 点であった。全国平均が 60 点であったとき、この学校の生徒の平均点は全国平均を下回っているといえるか。有意水準 5% で検定しなさい。



この場合

対立仮説 H_1 : 全国平均を下回っている 帰無仮説 H_0 : 全国平均と同じである
 となります。ここでは、全国平均を大きく下回っているときに仮説が棄却されるように片側検定を行います。

なお、今回は母標準偏差が分かっていませんが、標本の大きさが十分に大きいので、標本標準偏差を代わりに用います。

T 高校の母平均を m とする。

このとき、全国平均を下回っているならば $m < 60$ となる。

ここで、全国平均と同じ、つまり $m = 60$ であるという仮説を立てる。

標本の大きさは十分に大きいので、この仮説の下で、

標本平均 $\bar{X}$ は正規分布 $N\left(60, \frac{15^2}{100}\right)$ に従う。

これより、 $Z = \frac{\bar{X} - 60}{\frac{15}{10}}$ は近似的に標準正規分布 $N(0, 1)$ に従う。

正規分布表より $P(Z \geq -1.64) \approx 0.5 + 0.45 = 0.95$ なので、

有意水準 5% の棄却域は $Z \leq -1.64$ となる。

$\bar{X} = 57$ のとき、 $Z = \frac{57 - 60}{\frac{15}{10}} = -2$ であり、この値は棄却域に入るから仮説を棄却できる。

したがって、 $m < 60$ つまり全国平均を下回っていると判断してよい。

【中心極限定理】

母平均 m 、母標準偏差 σ の母集団から大きさ n の標本を無作為に抽出するとき、標本平均 $\bar{X}$ は、 n が大きいとき近似的に正規分布 $N\left(m, \frac{\sigma^2}{n}\right)$ に従う。

MEMO