統計学A

千葉大学理学部数学・情報数理学科
種村 秀紀
e-mail: tanemura@math.s.chiba-u.ac.jp
http://www.math.s.chiba-u.ac.jp/ ~ tanemura/index.html

平成 27 年 4 月 15 日

目 次

1 記述統計

母集団 : 調査したい対象
標本 : 母集団の中から選んだグループ.(母集団に比べ数は非常に少ない)
変数 : 調査項目として選ばれるある 1 つの特性
標本データ : 測定値, 変数の調査結果

- **変数の種類**

質的変数　形, 色, 性別 (数でないもの)
量的変数 { 離散型変数　(個数, 人数, 日数)
　　連続型変数　(時間, 身長, 重さ)

1.1 度数分布表

- **標本データ**

千葉大学男子 70 人の身長データ

169	171	172	167	171	176	164	169	168	164
169	168	164	159	161	167	162	174	168	165
169	167	165	171	168	170	163	177	162	164
172	177	175	173	156	163	159	157	172	174
182	161	175	170	175	173	167	154	173	168
175	164	169	171	161	163	176	155	166	180
168	164	176	168	181	173	159	183	168	166

母集団　: 千葉大学の男子学生
変数　: 千葉大学の男子学生の身長
標本の大きさ　: $n = 70$

最大値 183cm, 最小値 154cm, 範囲 (最大値 − 最小値)= 183 − 154 = 29cm.
階級の幅 3cm, 階級の数 10 で度数分布表を作る.

	階級	階級値	度数 f	累積度数 F	相対度数 f/n	累積相対度数 F/n
1	153.5 〜 156.5	155	3	3	0.043	0.043
2	156.5 〜 159.5	158	4	7	0.057	0.100
3	159.5 〜 162.5	161	5	12	0.071	0.171
4	162.5 〜 165.5	164	11	23	0.157	0.329
5	165.5 〜 168.5	167	14	37	0.200	0.529
6	168.5 〜 171.5	170	11	48	0.157	0.687
7	171.5 〜 174.5	173	9	57	0.129	0.814
8	174.5 〜 177.5	176	9	66	0.129	0.943
9	177.5 〜 180.5	179	1	67	0.014	0.957
10	180.5 〜 183.5	182	3	70	0.043	1.000

度数分布はこれをグラフに表せば直感的にいっそう理解し易くなる. 連続型変数に対してはヒストグラム（柱状図形）と呼ばれるグラフが有効である. 測定値が階級の境界の値になる場合には, 慣例に従って, その測定値は小さいほうの階級に入れることにする.

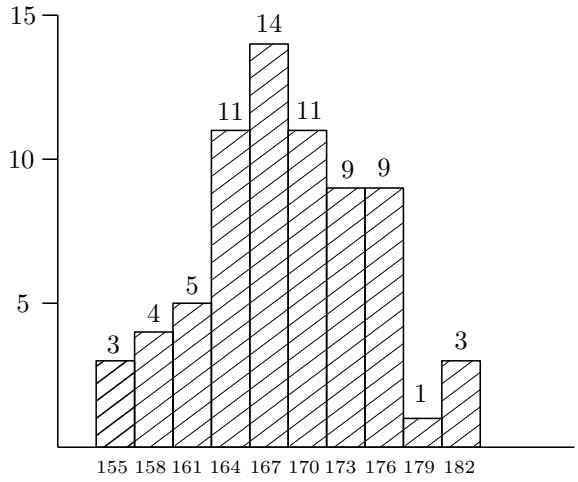


図 1. ヒストグラム（柱状グラフ）　千葉大学の学生（70 人）の身長

- 離散型変数についての度数分布表

  ある街で救急車が一日に何回出動したか記録がある (30 日間).
  変数 X: 救急車の一日における出勤回数, 標本の大きさ $n = 30$.

出動回数	x	0	1	2	3	4	5	6
度数	f	8	6	5	4	3	3	1
累積度数	F	8	14	19	23	26	29	30

離散型変数に対してもヒストグラムを用いることがあるが, ここでは棒グラフを用いる.

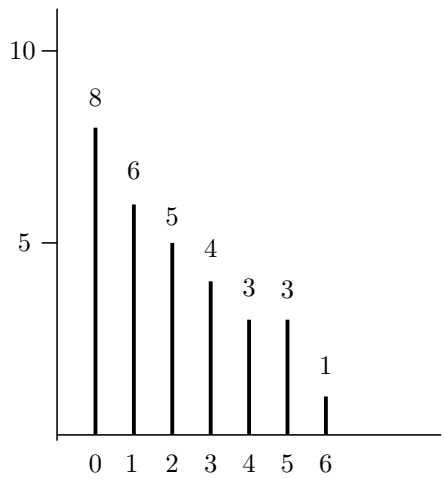


図 2. 棒グラフ　救急車の出動回数

1.2　算術的記述

位置の測度　分布の位置を代表させる特性値
　　(例) 平均, モード, メディアン
変動の測度　分布のばらつきを代表させる特性値
　　(例) 標準偏差, 四分位範囲

1.2.1 標本平均

定義 (標本平均,Sample mean)

(i) データ $x_1, x_2, \ldots, x_n$ が与えられているとき

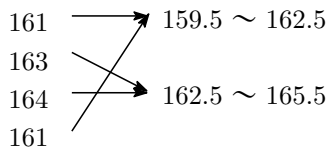
標本平均: $\overline{\mathbf{x}} \equiv \dfrac{1}{n}(x_1 + x_2 + \cdots + x_n) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$

(ii) 度数分表 $(y_i, f_i), i = 1, 2, \ldots, k,$ が与えられているとき

標本平均: $\overline{\mathbf{y}} \equiv \dfrac{1}{n}(y_1 f_1 + y_2 f_2 + \cdots + y_k f_k) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{k} y_i f_i$　k : 階級の数

注：$\equiv$ は右辺で左辺を定義するという意味で用いられている.

元のデータ



$$\frac{1}{4}(161 + 161 + 163 + 164) = 162.25$$

度数分布表

161	2
164	2

$$\frac{1}{4}(161 \times 2 + 164 \times 2) = 162.5$$

この例でもわかる様に元のデータを直接用いた標本平均と度数分布を用いた標本平均は一般に一致しない. これは階級の幅があるためで, 元のデータが異なるものでも階級値が同じになる場合があるからである. (上の例では 163, 164 が共に 164 の階級に属している) しかし, 標本の数 n が十分大きいときはこの誤差は無視できるほど小さくなる.

1.2.2 標本標準偏差

定義 (標本標準偏差, Sample standard deviation)

(i) データ $x_1, x_2, \ldots, x_n$ が与えられているとき

標本標準偏差　$s_x \equiv \sqrt{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \overline{\mathbf{x}})^2}$

(ii) 度数分表 $(y_i, f_i), i = 1, 2, \ldots, k,$ が与えられているとき

標本標準偏差　$s_y \equiv \sqrt{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{k}(y_i - \overline{\mathbf{y}})^2 f_i}$

標本標準偏差に関する計算の注意

度数分布表 (y_i, f_i) を用いたとき等式

$$s_y = \sqrt{\frac{1}{n-1}\left\{\left(\sum_{i=1}^{k} y_i^2 f_i\right) - \frac{1}{n}\left(\sum_{i=1}^{k} y_i f_i\right)^2\right\}}$$

が成り立つ. (実際の計算をするとき便利な形である.)

(証明)

和に関する性質：c と d が定数, つまり i に関係なく一定の時,

$$\sum_{i=1}^{k}(ca_i + db_i) = c\sum_{i=1}^{k} a_i + d\sum_{i=1}^{k} b_i \text{ (和の線形性という)}$$

度数の性質 : $\sum_{i=1}^{k} f_i = n$,
平均の定義 : $\overline{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{k} y_i f_i \Rightarrow \sum_{i=1}^{k} y_i f_i = n\overline{\mathbf{y}}$ を用いると,

$$(n-1)s_y^2 = \sum_{i=1}^{k}(y_i - \overline{\mathbf{y}})^2 f_i = \sum_{i=1}^{k}(y_i^2 - 2y_i\overline{\mathbf{y}} + \overline{\mathbf{y}}^2) f_i = \sum_{i=1}^{k} y_i^2 f_i - 2\overline{\mathbf{y}}\sum_{i=1}^{k} y_i f_i + \overline{\mathbf{y}}^2 \sum_{i=1}^{k} f_i$$
$$= \sum_{i=1}^{k} y_i^2 f_i - 2\frac{1}{n}\left(\sum_{i=1}^{k} y_i f_i\right)^2 + \frac{1}{n}\left(\sum_{i=1}^{k} y_i f_i\right)^2 = \sum_{i=1}^{k} y_i^2 f_i - \frac{1}{n}\left(\sum_{i=1}^{k} y_i f_i\right)^2$$

と計算できる. ■

(例) 千葉大学の学生 (70 人) の身長に対して標本平均と標本標準偏差を計算する.

	階級値 y_i	度数 f_i	$y_i f_i$	$y_i^2 f_i$
1	155	3	465	72705
2	158	4	632	99856
3	161	5	805	129605
4	164	11	1804	295856
5	167	14	2338	390446
6	170	11	1870	317900
7	173	9	1557	269361
8	176	9	1584	278784
9	179	1	179	32041
10	182	3	546	99372
計		70	11780	1985296

標本平均 $\overline{\mathbf{y}} = \frac{1}{70}\sum_{i=1}^{10} y_i f_i = \frac{11780}{70} \fallingdotseq \underline{168.3cm}$

分散=標本標準偏差の 2 乗 $s_y^2 = \frac{1}{70-1}\left\{\sum_{i=1}^{10} y_i^2 f_i - \frac{1}{70}(\sum_{i=1}^{10} y_i f_i)^2\right\}$
$= \frac{1}{69}\left\{1985296 - \frac{1}{70} \times (11780)^2\right\} = 41.88819884$

標本標準偏差 $s_y = \underline{6.5cm}$

注：有効桁を考慮して計算結果を書く事が大切である. ここでは有効桁を小数点以下一桁とした.

1.2.3　1 次変換を用いた計算

度数分布表　(y_i, f_i), $i = 1, 2, \ldots, k$ が与えられたとする.

1 次変換

$$\left.\begin{array}{l} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_k \end{array}\right\} \longrightarrow \left\{\begin{array}{l} z_1 = a_{11}y_1 + a_{12}y_2 + \cdots + a_{1k}y_k + b_1 \\ z_2 = a_{21}y_1 + a_{22}y_2 + \cdots + a_{2k}y_k + b_2 \\ \qquad\qquad \vdots \\ \qquad\qquad \vdots \\ z_k = a_{k1}y_1 + a_{k2}y_2 + \cdots + a_{kk}y_k + b_k \end{array}\right.$$

で定められる $y_1, \ldots, y_k$ から $z_1, \ldots, z_k$ への変換を **一次変換** という. この節で用いる 1 次変換は $a_{ij} = a$, $b_j = b$ と i, j に対して一定である場合,

$$y_i \to z_i = ay_i + b$$

に限る. このとき公式が成り立つ.

公式

度数分布表 (y_i, f_i) に対する標本平均を $\overline{\mathbf{y}}$, 標本標準偏差を s_y, 度数分布表 (z_i, f_i) に対する標本平均を $\overline{\mathbf{z}}$, 標本標準偏差を s_z とする.

$$(1)\quad \overline{\mathbf{z}} = a\overline{\mathbf{y}} + b$$
$$(2)\quad s_z = |a| s_y$$

証明

$$\begin{aligned}
(1)\quad \overline{\mathbf{z}} &= \frac{1}{n}\sum_{i=1}^{k} z_i f_i = \frac{1}{n}\sum_{i=1}^{k}(ay_i + b)f_i \\
&= \frac{a}{n}\sum_{i=1}^{k} y_i f_i + b\frac{1}{n}\sum_{i=1}^{k} f_i = a\overline{\mathbf{y}} + b \\
(2)\quad s_z^2 &= \frac{1}{n-1}\sum_{i=1}^{k}(z_i - \overline{\mathbf{z}})^2 f_i \\
&= \frac{1}{n-1}\sum_{i=1}^{k}(ay_i + b - a\overline{\mathbf{y}} - b)^2 f_i \\
&= a^2\frac{1}{n-1}\sum_{i=1}^{k}(y_i - \overline{\mathbf{y}})^2 f_i = a^2 s_y^2 \qquad \blacksquare
\end{aligned}$$

この公式を用いると, 1 次変換を用いて標本平均, 標本標準偏差の計算を簡単に行うことができる.

$$\begin{array}{ccc}
(y_i, f_i) & \longrightarrow & (z_i, f_i) \\
 & \text{1 次変換} & \\
\downarrow & & \downarrow \\
\text{標本平均 } \overline{\mathbf{y}} & & \text{標本平均 } \overline{\mathbf{z}} \\
\text{標本標準偏差 } s_y & \longleftarrow & \text{標本標準偏差 } s_z \\
 & \text{1 次変換} &
\end{array}$$

(例) 千葉大学の学生 (70 人) の身長に対して 1 次変換を用いて標本平均, 標本標準偏差を計算してみる. 度数は 167 の時最大であって, 階級の幅が 3cm であるという理由から,

$$z_i = \frac{y_i}{3} - \frac{167}{3},\ i = 1, 2, \ldots, k$$

つまり $a = \frac{1}{3}, b = -\frac{167}{3}$ という一次変換を用いる.

	階級値 z_i	度数 f_i	$z_i f_i$	$z_i^2 f_i$
1	-4	3	-12	48
2	-3	4	-12	36
3	-2	5	-10	20
4	-1	11	-11	11
5	0	14	0	0
6	1	11	11	11
7	2	9	18	36
8	3	9	27	81
9	4	1	4	16
10	5	3	15	75
計		70	30	334

まず (z_i, f_i) に対して標本平均と標本標準偏差を計算する.

$$\begin{aligned}
\overline{\mathbf{z}} &= \frac{1}{70}\sum_{i=1}^{10} z_i f_i = \frac{30}{70} = 0.428571423 \\
s_z^2 &= \frac{1}{69}(334 - \frac{1}{70} \times 30^2) = 4.654244306 \\
s_z &= 2.157369765
\end{aligned}$$

次に 1 次変換 $y_i = 3z_i + 167,\ i = 1, 2, \ldots, k$ に対して公式を用いると,

$$\begin{aligned}
\overline{\mathbf{y}} &= 3\overline{\mathbf{z}} + 167 \fallingdotseq \underline{168.3cm} \\
s_y &= 3s_z \fallingdotseq \underline{6.5cm}
\end{aligned}$$

となり, 直接計算した場合と同じ値が得られることが確認できた.

1.3 その他の記述的測度

1.3.1 位置の測度

- **最頻値, モード (Mode)**

> **定義** (最頻値)
>
> 離散型変数 (または質的変数) が最大度数 f を持つ x が 1 つあるときその値 x を最頻値 (モード) と呼ぶ. 2 つ以上あるときは, 最頻値なしという.

(例) 救急車の 1 日における出動回数 (30 日間)

出勤回数	0	1	2	3	4	5	6
度数 f	8	6	5	4	3	3	1
累積度数 F	8	14	19	23	26	29	30

最大の度数を持つ測定値は x=0, モード 0 回

- **中央値, メディアン (Median)**

> **定義** 中央値 (離散型変数の場合)
>
> 離散変数データを大きさの順に並べたとき, 標本の数 n が奇数である場合 $\frac{n+1}{2}$ 番目の測定値, 偶数である場合 $\frac{n}{2}$ 番目と $\frac{n}{2}+1$ 番目の測定値の算術平均を中央値という.

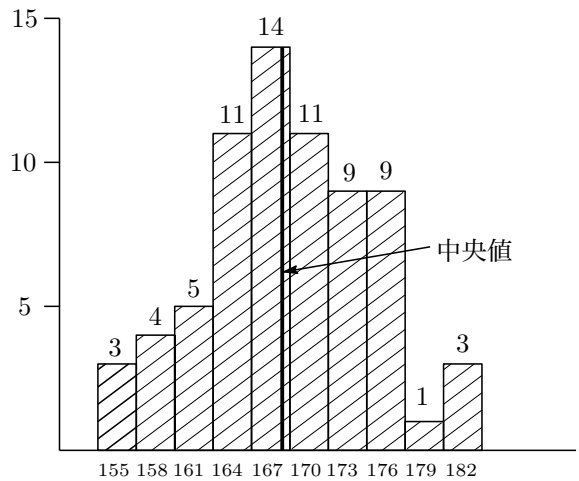
(例) 救急車の出動回数の例では $n = 30$ の偶数であるから 15 番目と 16 番目のデータでの算術平均

15 番目の値 2 回, 16 番目の値 2 回　$\because$　メディアン=2 回

> **定義** 中央値 (連続型変数の場合)
>
> 柱状グラフ（ヒストグラム) の面積を半分にする値を中央値と呼ぶ.

(例) 千葉大学 70 人の身長の例では, 標本数 $n = 70$, 階級の幅 $3cm$ ヒストグラムの面積は $3 \times 70 = 210$. 従って半分の面積は $210 \times \frac{1}{2} = 105$.

階級	度数 f	累積度数 F	累積面積
1	3	3	9
2	4	7	21
3	5	12	36
4	11	23	69
5	14	37	111
6	11	48	144
7	9	57	171
8	9	66	198
9	1	67	201
10	3	70	210



累積面積より中央値 $\mathbf{y}_M$ は階級 5(165.5 以上 168.5 未満) に含まれており,

$$69 + 14(\mathbf{y}_M - 165.5) = 105$$

を満たす事がわかる. ゆえに,

$$\mathbf{y}_M = 165.5 + \frac{105 - 69}{14} = 165.5 + \frac{36}{14} \fallingdotseq 168.1cm$$

つまり中央値は 168.1cm である.

（標本平均と中央値)：上の例でもわかるように, 中央値 168.1 と標本平均 168.3 は一般には一致しない. 典型的な例を挙げておく.

データ 1

点	30	60	90
度数	5	10	5

標本平均 60 点, 中央値 60 点

データ 2

点	0	60	90
度数	5	5	10

標本平均 60 点, 中央値 75 点

1.3.2 変動の測度

- **範囲, レンジ (Range)**

定義 範囲

元のデータ $x_1, x_2, \ldots, x_n$ を用いる.

$$\text{範囲} \equiv (\text{データの最大値}) - (\text{データの最小値})$$

- **四分位範囲 (Interquaritile range)**

定義 四分位範囲

$$\text{四分位範囲} \equiv (\text{第 3 四分位数}) - (\text{第 1 四分位数})$$

四分位範囲の定義になる第 1 四分位数と第 3 四分位数の定義を述べる. 第 2 四分位数は中央値である.

定義 (離散型変数の場合)

離散データを小さい順にならべたとき標本の数 n が

$$4 \text{ の倍数のとき} \quad \Rightarrow \quad \begin{cases} \dfrac{n}{4} \text{番目と} \dfrac{n}{4}+1 \text{ 番目の値の算術平均を} \\ \text{第 1 四分位数と呼ぶ.} \\ \dfrac{3n}{4} \text{番目と} \dfrac{3n}{4}+1 \text{ 番目の値の算術平均を} \\ \text{第 3 四分位数と呼ぶ.} \end{cases}$$

$$4 \text{ の倍数でないとき} \quad \Rightarrow \quad \begin{cases} \left[\dfrac{n}{4}\right]+1 \text{ 番目の値を第 1 四分位数と呼ぶ.} \\ \\ \left[\dfrac{3n}{4}\right]+1 \text{ 番目の値を第 3 四分位数と呼ぶ.} \end{cases}$$

ここで $\left[\dfrac{n}{4}\right]$ は $\dfrac{n}{4}$ を超えない最大の整数.

(例) 救急車の出動回数の場合

$$\begin{aligned} &\text{第 1 四分位数} \quad = 8 \text{ 番目のデータ=0 回} \\ &\text{第 3 四分位数} \quad = 23 \text{ 番目のデータ=3 回} \\ &\text{四分位範囲} \quad\ \ = 3 \text{ 回} - 0 \text{ 回=3 回} \end{aligned}$$

階級	度数 f	累積度数 F
0	8	8
1	6	14
2	5	19
3	4	23
4	3	26
5	3	29
6	1	30

定義 (連続型変数の場合)

柱状グラフ (ヒストグラム) の面積を 4 等分する値 $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3$ を順に第 1 四分位数, 第 2 四分位数 (=中央値), 第 3 四分位数という.

(例) 千葉大学学生 70 人の身長のデータに対する四分位数と四分位範囲の計算.

- **第 1 四分位数**

  $\mathbf{y}_1$ は左側の面積を $\dfrac{210}{4} = 52.5$ にする値であり, 次の等式を満足

$$\begin{aligned} 36 + (\mathbf{y}_1 - 162.5) \times 11 &= 52.5 \\ \mathbf{y}_1 = 162.5 + \frac{52.5 - 36}{11} \\ &= 164.0cm \end{aligned}$$

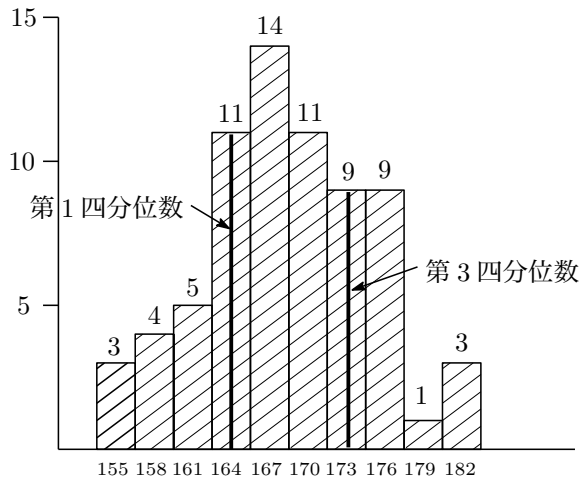
- **第 3 四分位数**

  x_3 は左側の面積を $\dfrac{3}{4} \times 210 = 157.5$ にする値であり次の資料を満足

$$\begin{aligned} 144 + (\mathbf{y}_3 - 171.5) \times 9 &= 157.5 \\ \mathbf{y}_3 = 171.5 + \frac{157.5 - 144}{9} &= 173cm \end{aligned}$$

  四分位範囲=$\mathbf{y}_3 - \mathbf{y}_1 = \underline{9cm}$

階級	度数 f	累積度数 F	累積面積
1	3	3	9
2	4	7	21
3	5	12	36
4	11	23	69
5	14	37	111
6	11	48	144
7	9	57	171
8	9	66	198
9	1	67	201
10	3	70	210



2 確率

- **標本空間** (Sample Space)

  試行で得られる結果すべての集まり. 記号は S を用いる.

- **標本点** (Sample point)

  標本空間に含まれる要素を標本点という.

- **事象** (Event)

  S の部分集合を事象という. A, B, C などの記号を用いる. つぎの事象は特別例である.

  空事象: 標本点を 1 つも含まない場合も事象であり空事象と呼び ϕ または $\{\}$ と書く.

  全事象: 全ての標本点を含む場合も事象であり全事象と呼び, 標本空間と同じであるので S と書く.

  根元事象: 1 つの標本点からなる事象を根元事象と呼び, 例えば i を 1 つの標本点とすると $\{i\}$ となる.

- **確率** (Probability)

  各事象に 0 以上 1 以下の数を対応させるもの.

2.1 試行と確率

- **試行 1**

  箱の中に 1〜6 の番号がついた玉が各々1 個ずつ, 計 6 個入っている. この中から無作為に玉を 1 個取り出し, その番号を調べる試行を行う. (無作為に取り出す = 各々の玉が取り出される確率が等しい.)

  標本空間 $S = \{1, 2, 3, 4, 5, 6\}$

  事象 空事象 ϕ, 全事象 S, 根元事象 $\{i\}$, $1 \leq i \leq 6$ のほか, 次の事象がある.
  (計 $2^6 = 64$ 個ある.)

2 つの標本点からなる事象	$\{i, j\}$	$1 \leq i < j \leq 6$
3 つの標本点からなる事象	$\{i, j, k\}$	$1 \leq i < j < k \leq 6$
4 つの標本点からなる事象	$\{i, j, k, \ell\}$	$1 \leq i < j < k < \ell \leq 6$
5 つの標本点からなる事象	$\{i, j, k, \ell, m\}$	$1 \leq i < j < k < \ell < m \leq 6$

  確率 根元事象の確率は $\frac{1}{6}$, つまり, $P(\{i\}) = \frac{1}{6}, 1 \leq i \leq 6$
  A が k 個の標本点から構成されている事象のとき $P(A) = \frac{k}{6}$

- **試行 2**

  箱の中に赤球が 3 個, 青球が 2 個, 緑球が 1 個計 6 個入っている. この中から無作為に玉を 1 個取り出し, その玉の色を調べる試行を行う.

  標本空間 $S = \{$ 赤, 青, 緑 $\}$

  事象 $\phi, \{$ 赤 $\}, \{$ 青 $\}, \{$ 緑 $\}, \{$ 赤, 青 $\}, \{$ 赤, 緑 $\}, \{$ 青, 緑 $\}, S$

  確率

  $$\begin{cases} P(\phi) & = 0 \\ P(\{\text{赤}\}) & = \frac{3}{6} = \frac{1}{2} \\ P(\{\text{青}\}) & = \frac{2}{6} = \frac{1}{3} \\ P(\{\text{緑}\}) & = \frac{1}{6} \\ P(\{\text{赤, 青}\}) & = \frac{1}{2} + \frac{1}{3} = \frac{5}{6} \\ P(\{\text{赤, 緑}\}) & = \frac{1}{2} + \frac{1}{6} = \frac{2}{3} \\ P(\{\text{青, 緑}\}) & = \frac{1}{3} + \frac{1}{6} = \frac{1}{2} \\ P(S) & = 1 \end{cases}$$

試行 1 と 2 の対応

計 6 個の玉のうち 1, 2, 3 の番号が付いているものは赤玉, 4, 5 の番号が付いているものは青玉, 6 の番号が付いているものは緑玉とすると,

$$\begin{array}{lll} P(\{1,2,3\}) & = \frac{1}{2} & = P(\{\text{ 赤 }\}) \\ P(\{4,5\}) & = \frac{1}{3} & = P(\{\text{ 青 }\}) \\ P(\{6\}) & = \frac{1}{6} & = P(\{\text{ 緑 }\}) \end{array}$$

が成り立つことがわかる.

- **試行 3**(復元抽出法)

箱の中に 1〜6 の番号のついた玉が計 6 個入っている. この中から無作為に玉を 1 個取り出し, その番号を調べ元の箱に戻す. そして同じ箱の中から無作為にもう一度玉を取り出しその番号を調べるという試行を行う.

標本空間: 1 回目に i, 2 回目に j を取り出したという標本点を (i,j) と書く. すると,

$$S = \left\{ \begin{array}{cccccc} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{array} \right\}$$

$= \{(i,j); \quad 1 \leq i,j \leq 6\}$

標本空間 S に含まれる標本点の総数 $\#S = 36$

事象: 空事象, 全事象, 根元事象など計 2^{36} 個ある.

確率: 根元事象の確率は $\frac{1}{36}$, つまり, $P(\{(i,j)\}) = \frac{1}{36}, \quad (i,j) \in S$ であり,
A が k 個の標本点から構成されている事象のとき $P(A) = \frac{k}{36}$

- **試行 4** (非復元抽出法)

箱の中に 1〜6 の番号のついた玉が計 6 個入っている. この中から無作為に玉を 1 個取り出す (ただし取り出した玉は箱に戻さない). そして同じ箱の中から無作為にもう一度玉を取り出しその番号を調べるという試行を行う.

標本空間: 1 回目に i, 2 回目に j を取り出したという標本点を (i,j) と書く. すると,

$$S = \left\{ \begin{array}{cccccc} & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & \end{array} \right\}$$

$= \{(i,j); \quad 1 \leq i,j \leq 6, \quad i \neq j\}$

標本空間 S に含まれる標本点の総数 $\#S = 30$

事象: 空事象, 全事象, 根元事象など計 2^{30} 個ある.

確率: 根元事象の確率は $\frac{1}{30}$, つまり, $P(\{(i,j)\}) = \frac{1}{30}, \quad (i,j) \in S$ であり,
A が k 個の標本点から構成されている事象のとき $P(A) = \frac{k}{30}$

一般化について

玉の数を N, 取り出す回数を n とする. 試行 3, 4 の一般化を考えてみる.

(試行 3)

$N=2, n=5$　表がでる確率が $\frac{1}{2}$, 裏が出る確率が $\frac{1}{2}$ であるコインを 5 回投げて結果を順に記録する.

$N=6, n=10$　各目が出る確率が全て等しく $\frac{1}{6}$ であるサイコロを 10 回投げて結果を順に記録する.

$N=38, n=4$　ルーレットを 4 回まわしてその結果を順に記録する.

(試行 4)

$N=52, n=5$　トランプから無作為にカードを順に 5 枚取り出し, 結果を記録する.

$N=1000, n=3$　ある学校の生徒 1000 人から無作為に順に 3 人選び, 生徒会長, 生徒会副会長, 書記を決める.

試行 3 と 4 の確率の差に関する注意 ($n=2$ の場合)

(試行 3)　$P(\{i,j\})=\frac{1}{N^2}$

(試行 4)　$P(\{i,j\})=\begin{cases}\frac{1}{N(N-1)} & (i\neq j)\\ 0 & (i=j)\end{cases}$

２つのモデルの根元事象の確率の差は

$$\frac{1}{N(N-1)}-\frac{1}{N^2}=\frac{N-(N-1)}{N^2(N-1)}=\frac{1}{N^2(N-1)}$$

例えば $N=100$ のとき,

確率は　1 万分の 1 程度
差は　100 万分の 1 程度
$\Longrightarrow$ モデル 3 と 4 は差があるが N が大きいときには無視できるくらい小さい

- **試行 5**

箱の中に 1〜6 の番号がついた玉が各々1 個ずつ計 6 個入っている. この中から一度に 2 個の玉を無作為に取り出す試行を行った.

標本空間: 番号 i と番号 j の玉を取り出すという標本点を $[i,j]$ と書く $(i<j)$. すると,

$$S=\left\{\begin{array}{ccccc}[1,2] & [1,3] & [1,4] & [1,5] & [1,6]\\ & [2,3] & [2,4] & [2,5] & [2,6]\\ & & [3,4] & [3,5] & [3,6]\\ & & & [4,5] & [4,6]\\ & & & & [5,6]\end{array}\right\}$$

$=\{[i,j];\quad 1\leq i<j\leq 6\}$

標本空間 S に含まれる標本点の総数 $\#S=15$

事象: 空事象, 全事象, 根元事象など計 2^{15} 個ある.

確率: 根元事象の確率は $\frac{1}{15}$, つまり, $P(\{(i,j)\})=\frac{1}{15}$,　$[i,j]\in S$ であり, A が k 個の標本点から構成されている事象のとき $P(A)=\frac{k}{15}$

試行 4 と 5 の対応

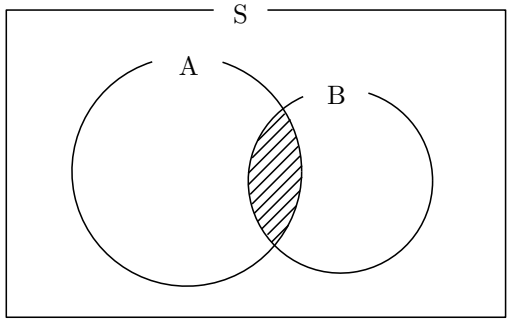
試行 5 で i 番と j 番の玉を一度に取り出す確率は, 試行 4 で 1 回目に i 番, 2 回目に j 番の玉を取り出す確率と, 1 回目に j 番, 2 回目に i 番の玉を取り出す場合を合わせた確率と同じになる.

つまり, $P(\{(i,j),(j,i)\})=P(\{[i,j]\}),\quad 1\leq i<j\leq 6$ が成り立つ.

2.2 条件付確率と独立性

A と B を事象とする. A にも B にも含まれている標本点からなる事象を A と B の積事象と呼び $A \cap B$ と書く.

A かまたは B に含まれている標本点からなる事象を A と B の和事象と呼び $A \cup B$ と書く.



$A \cap B = \phi$ が成り立つとき A と B は**互いに素** (または排反) であるという. A と B が互いに素であるとき,

$$P(A \cup B) = P(A) + P(B)$$

が成り立つ. この式を**加法の公式**という.

定義 (条件付確率)

事象 B が起こるという条件の下での事象 A が起こる条件付確率を

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

で定義する. ただし $P(B) = 0$ の時は定義しない.

定義 (独立性)

事象 A と事象 B が**独立**であるとは

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

が成り立つ事である.

$P(B) \neq 0$ の時 A と B が独立 $\Longleftrightarrow P(A|B) = P(A)$ であることが分かる.

また,

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

が成り立つ. この式を**乗法の公式**という.

例) 試行 3 (復元抽出法) を考える.

1 回目に取り出した玉が i 番である事象を $A(i)$ と書く. つまり,

$$A(i) = \{(i,1),(i,2),(i,3),(i,4),(i,5),(i,6)\}$$

2 回目に取り出した玉が j 番である事象を $B(j)$ と書く. つまり,

$$B(j) = \{(1,j),(2,j),(3,j),(4,j),(5,j),(6,j)\}$$

$A(i), B(j)$ とその積事象を計算してみると,

$$P(A(i) \cap B(j)) = P(\{(i,j)\}) = \frac{1}{36}$$
$$P(A(i)) = \frac{1}{6}, \quad P(B(j)) = \frac{1}{6}$$

となるので,

$$P(A(i) \cap B(j)) = P(A(i))P(B(j))$$

及びその同値の式

$$P(A(i)) = P(A(i)|B(j))$$

が成り立つ. 従って $A(i)$ と $B(j)$ は独立である.

例) 試行 4 (非復元抽出法) を考える. この場合,

$$A(i) = \{(i,j); 1 \leq j \leq 6, \quad i \neq j\}$$
$$B(j) = \{(i,j); 1 \leq i \leq 6, \quad i \neq j\}$$

となるので $A(i), B(j)$ とその積事象の確率を計算してみると,

$$P(A(i)) = \sum_{\substack{j=1 \\ j \neq i}}^{6} P(\{(i,j)\}) = \frac{5}{30} = \frac{1}{6}$$

$$P(B(j)) = \sum_{\substack{i=1 \\ i \neq j}}^{6} P(\{(i,j)\}) = \frac{5}{30} = \frac{1}{6}$$

$$P(A(i) \cap B(j)) = P(\{(i,j)\}) = \begin{cases} \frac{1}{30} & i \neq j \\ 0 & i = j \end{cases}$$

となり,

$$P(A(i) \cap B(j)) \neq P(A(i))P(B(i))$$
$$P(A(i)|B(j)) = \begin{cases} \frac{\frac{1}{30}}{\frac{1}{6}} = \frac{1}{5} & i \neq j \\ 0 & i = j \end{cases}$$

が成り立つ. したがって $A(i)$ と $B(j)$ は独立ではない.

例) 試行 3 (復元抽出法) を考える.

1 回目の取り出しで偶数番の玉を得たという事象を C と書く. つまり,

$$C = \begin{Bmatrix} (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{Bmatrix}$$

1 回目と 2 回目で取り出した玉の番号の和は 7 以上である事象を D と書く. つまり,

$$D = \begin{Bmatrix} (1,6) & (2,5) & (2,6) & (3,4) & (3,5) & (3,6) \\ (4,3) & (4,4) & (4,5) & (4,6) & (5,2) & (5,3) \\ (5,4) & (5,5) & (5,6) & (6,1) & (6,2) & (6,3) \\ (6,4) & (6,5) & (6,6) & & & \end{Bmatrix}$$

C, D の確率は,

$$P(C) = \frac{18}{36} = \frac{1}{2}, \quad P(D) = \frac{21}{36} = \frac{7}{12}$$

となり, その積事象

$$C \cap D = \left\{ \begin{array}{cccccc} (2,5) & (2,6) & (4,3) & (4,4) & (4,5) & (4,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{array} \right\}$$

の確率は,

$$P(C \cap D) = \frac{12}{36} = \frac{1}{3}$$

が成り立つ. 従って,

$$P(C \cap D) = \frac{1}{3} \neq \frac{1}{2} \times \frac{7}{12} = P(C)P(D)$$

となり, C と D は独立ではない.

2.3 ベイズの定理 (条件付確率での原因と結果の考察)

例 病気の検査の精度

まれにしか起こらない病気を発見するのに有効な検査法がある. この検査法では,

実際に病気の人	$\Rightarrow$	97% の確率で陽性
健康な人	$\Rightarrow$	5% の確率で陽性
風邪の人	$\Rightarrow$	10% の確率で陽性

となることが今までのデータでわかっている. 一方, 多人数からなるある母集団において

実際に病気の人の割合が 1%
健康な人の割合が 96%
風邪の人の割合が 3%

であることがわかっている. この母集団から無作為に選ばれた一人が陽性反応を示した. このときこの人が実際に病気である確率はどのくらいか調べる.

実際に病気である事象を**病**
健康である事象を**健**
風邪である事象を**風**
陽性反応を示す事象を**陽**

と書くことにする. この記号を用いて上で述べたことをまとめると,

$$\begin{array}{llll} \mathrm{P}(\text{陽} \mid \text{病}) & = 0.97 & \mathrm{P}(\text{病}) & = 0.01 \\ \mathrm{P}(\text{陽} \mid \text{健}) & = 0.05 & \mathrm{P}(\text{健}) & = 0.96 \\ \mathrm{P}(\text{陽} \mid \text{風}) & = 0.10 & \mathrm{P}(\text{風}) & = 0.03 \end{array}$$

となり, 求める確率は $\mathrm{P}(\text{病} \mid \text{陽})$ である. 乗法の公式より,

$$\begin{array}{llll} P(\text{病} \cap \text{陽}) & = P(\text{病})P(\text{陽} \mid \text{病}) & = 0.01 \times 0.97 & = 0.0097 \\ P(\text{健} \cap \text{陽}) & = P(\text{健})P(\text{陽} \mid \text{健}) & = 0.96 \times 0.05 & = 0.048 \\ P(\text{風} \cap \text{陽}) & = P(\text{風})P(\text{陽} \mid \text{風}) & = 0.03 \times 0.1 & = 0.003 \end{array}$$

が成り立ち, 加法の公式より,

$$\begin{aligned} P(\text{陽}) &= P(\text{病} \cap \text{陽}) + P(\text{健} \cap \text{陽}) + P(\text{風} \cap \text{陽}) \\ &= 0.0097 + 0.048 + 0.003 \\ &= 0.0607 \end{aligned}$$

が成り立つ. 従って,

$$P(\text{病} \mid \text{陽}) = \frac{P(\text{病} \cap \text{陽})}{P(\text{陽})} = \frac{0.0097}{0.0607} = 0.16$$

従って検査で陽性であっても本当に病気である確率は 16% 程度である.

次に逆の場合を考えてみる. 無作為に選ばれた一人が陰性反応を示したとき, この人が実際に病気である確率 $P(\text{病} \mid \text{陰})$ である確率を調べる.

$$\begin{aligned} \mathrm{P}(\text{陰} \mid \text{病}) \quad &= 0.03, \quad \mathrm{P}(\text{陰} \mid \text{健}) = 0.95, \quad \mathrm{P}(\text{陰} \mid \text{風}) = 0.9 \\ P(\text{病} \mid \text{陰}) \quad &= \tfrac{P(\text{病} \cap \text{陰})}{P(\text{陰})} \\ P(\text{病} \cap \text{陰}) \quad &= P(\text{病})P(\text{陰} \mid \text{病}) = 0.01 \times 0.03 = 0.0003 \\ P(\text{陰}) \quad &= 1 - P(\text{陽}) = 1 - 0.0607 = 0.9393 \end{aligned}$$

従って,

$$P(\text{病} \mid \text{陰}) = \frac{0.0003}{0.9393} = 0.000319 \quad (0.0319\%)$$

となり, 陰性であれば本当に病気である確率は殆ど無いことが分かる.

上の例は次で示されるベイズの定理の応用例である.

ベイズの定理

$A_1, A_2, \ldots, A_n$ を互いに素である事象列, つまり,

$$A_i \cap A_j = \phi, \quad i \neq j$$

であるとし, さらに, $\cup_{i=1}^{n} A_i = S$ を満たすものとする. 各 i に対して $P(A_i) > 0$ であり, B も $P(B) > 0$ である事象とすると

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

が成り立つ.

証明　乗法の公式より,

$$\begin{aligned} P(B|A_k)P(A_k) \quad &= P(B \cap A_k) \\ \textstyle\sum_{i=1}^{n} P(B|A_i)P(A_i) \quad &= \textstyle\sum_{i=1}^{n} P(B \cap A_i) \end{aligned}$$

が成り立つ. $\cup_{i=1}^{n} A_i = S$ であり, $A_i \cap A_j = \phi$, $i \neq j$ であることより, 加法の公式を用いると,

$$\sum_{i=1}^{n} P(B \cap A_i) = P(B \cap S) = P(B)$$

となる. 従って,

$$\frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)} = \frac{P(B \cap A_k)}{P(B)} = P(A_k|B)$$

が得られる.

2.4 確率分布

2.4.1 離散確率分布

標本空間が有限個の標本点から構成される場合, つまり $S = \{x_1, x_2, \ldots, x_n\}$ の場合を考える. ここで n は自然数である.

定義 (確率密度) p が S 上の確率密度であるとは,

(i) $0 \leq p(x) \leq 1, \quad x \in S$

(ii) $\displaystyle\sum_{x \in S} p(x) = p(x_1) + p(x_2) + \cdots + p(x_n) = 1$

が成り立つことである.

S の事象 $A = \{y_1, y_2, \ldots, y_k\}$ に対して,

$$P(A) = \sum_{x \in A} p(x) = p(y_1) + p(y_2) + \cdots + p(y_k),$$

で P を定義すると (i) $0 \leq P(A) \leq 1$, (ii) $P(S) = 1$ が成立し, 加法の公式を満足する. 従って P は S 上の確率である.

以後 S の元, つまり標本点は実数とする (ほとんどの例は整数).

平均, 2 次モーメント, 分散, 標準偏差は次で定義される.

(i) **平均 (mean)** $\mu = \displaystyle\sum_{x \in S} xp(x)$

(ii) **2 次モーメント (second moment)** $m_2 = \displaystyle\sum_{x \in S} x^2 p(x)$

(iii) **分散 (variance)** $V = \displaystyle\sum_{x \in S} (x - \mu)^2 p(x) = m_2 - \mu^2$

(iv) **標準偏差 (standard deviation)** $\sigma = \sqrt{V}$

2.4.2 離散確率分布の例

(i) **ベルヌーイ分布 (Bernoulli distribution)**

$$S = \{0, 1\}, \quad p(0) = 1 - p, \quad p(1) = p,$$

(ただし p は 0 以上 1 以下の実数) の時, 確率密度 $p(x)$ で定まる (離散) 確率分布をベルヌーイ分布という.

- **対応する試行**
  表が出る確率が p, 裏が出る確率が $1 - p$ であるコインを 1 回投げて, 表が出た回数を調べる.
- **平均** $\mu = 0 \cdot p(0) + 1 \cdot p(1) = p$
- **2 次モーメント** $m_2 = 0^2 \cdot p(0) + 1^2 \cdot p(1) = p$
- **分散** $V = m_2 - \mu^2 = p - p^2 = p(1 - p)$
- **標準偏差** $\sigma = \sqrt{V} = \sqrt{p(1 - p)}$

(ii) **離散一様分布 (Uniform distrbution)**

$$S = \{1, 2, \ldots, n\}, \quad p(x) = \tfrac{1}{n}, \quad x \in S$$

の時, 確率密度 $p(x)$ で定まる (離散) 確率分布を離散一様分布という.

- **対応する試行**
  各々の目が出る確率がすべて等しいサイコロを投げ, 出た目の数を調べる. (この場合 $n = 6$)
- **平均** $\mu = \displaystyle\sum_{x=1}^{n} xp(x) = \frac{1}{n}\sum_{x=1}^{n} x = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$
- **2 次モーメント** $m_2 = \displaystyle\sum_{x=1}^{n} x^2 p(x) = \frac{1}{n}\sum_{x=1}^{n} x^2 = \frac{1}{n}\frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}$
- **分散** $V = m_2 - \mu^2 = \dfrac{(n+1)(2n+1)}{6} - \left(\dfrac{n+1}{2}\right)^2 = \dfrac{n+1}{12}\{2(2n+1) - 3(n+1)\} = \dfrac{n^2-1}{12}$
- **標準偏差** $\sigma = \sqrt{V} = \sqrt{\dfrac{n^2-1}{12}}$

(iii) **二項分布 (Binomial distribution)**

$$S = \{0, 1, 2, \ldots, n\}, \quad p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x \in S$$

の時, 確率密度 $p(x)$ で定まる確率分布を母数 (n, p) の二項分布という. ($B_i(n, p)$ とかく.) ここで, $\binom{n}{x}$ は n 個の異なるものから x 個を選ぶ組み合わせの数であり, n の階乗 $n! = n \cdot (n-1) \cdots 3 \cdot 2 \cdot 1$ (ただし, $0! = 1$ と定義) を用いて

$$\binom{n}{x} = \frac{n!}{(n-x)!x!}, \quad n \text{ は自然数, } x \text{ は } 0 \text{ 以上 } n \text{ 以下の整数}$$

と表すことができる. $\binom{n}{x}$ は漸化式

$$\binom{n+1}{x} = \binom{n}{x} + \binom{n}{x-1}, \quad n \text{ は自然数, } x \text{ は } 0 \text{ 以上 } n \text{ 以下の整数}$$

をみたすが, これは $n+1$ 個の異なるものから x 個を選ぶ時, 特別に定めた 1 個を選ぶ場合と選ばない場合に分けて数えあげたことに対応している.

- 対応する試行
  表が出る確率が p, 裏が出る確率 $1-p$ であるコインを n 回投げて表が出た回数を調べる.
  表が出た回数を X とおき, $n = 1, 2, 3$ の時に確率分布を調べてみる.

  $n = 1$ の時 (ベルヌーイ分布と一致)

  $X = 0 \iff$ 裏 $\cdots$ $p(0) = 1 - p$
  $X = 1 \iff$ 表 $\cdots$ $p(1) = p$

  $n = 2$ の時

  $X = 0 \iff$ 裏 · 裏 $\cdots$ $p(0) = (1-p)^2$
  $X = 1 \iff$ 表 · 裏または裏 · 表 $\cdots$ $p(1) = 2p(1-p)$
  $X = 2 \iff$ 表 · 表 $\cdots$ $p(2) = p^2$

  $n = 3$ の時

  $X = 0 \iff$ 裏 · 裏 · 裏 $\cdots$ $p(0) = (1-p)^3$
  $X = 1 \iff$ 裏 · 裏 · 表, 裏 · 表 · 裏または表 · 裏 · 裏 $\cdots$ $p(1) = 3p(1-p)^2$
  $X = 2 \iff$ 裏 · 表 · 表, 表 · 裏 · 表 または 表 · 表 · 裏 $\cdots$ $p(2) = 3p^2(1-p)$
  $X = 3 \iff$ 表 · 表 · 表 $\cdots$ $p(3) = p^3$

演習　$n=4$ の時, $p(0),p(1),p(2),p(3),p(4)$ を求めよ.

参照として特別な場合の二項分布の確率を記しておく.

$n=10,p=0.5$ のとき		$n=10,p=0.01$ のとき	
$p(0)=0.0010$	$p(6)=0.2051$	$p(0)=0.9044$	$p(6)=0$
$p(1)=0.0098$	$p(7)=0.1172$	$p(1)=0.0914$	$p(7)=0$
$p(2)=0.0439$	$p(8)=0.0439$	$p(2)=0.0042$	$p(8)=0$
$p(3)=0.1172$	$p(9)=0.0098$	$p(3)=0.0001$	$p(9)=0$
$p(4)=0.2051$	$p(10)=0.0010$	$p(4)=0$	$p(10)=0$
$p(5)=0.2461$		$p(5)=0$	

さらに一般の場合は二項分布表を見よ. 二項分布の平均, 分散を計算するには二項定理を用いる.

二項定理： $(p+q)^n=\sum_{x=0}^{n}\binom{n}{x}p^xq^{n-x}$

$q=1-p$ とおく. この二項定理を用いると, $P(S)=1$ が直ちに導かれる:

$$P(S)=\sum_{x=0}^{n}p(x)=\sum_{x=0}^{n}\binom{n}{x}p^xq^{n-x}=(p+q)^n=1.$$

二項分布の平均と分散の計算

$$\mu=\sum_{k=0}^{n}k\cdot p(k)=\sum_{k=0}^{n}k\times\frac{n!}{k!(n-k)!}p^kq^{n-k}=\sum_{k=1}^{n}\frac{n!}{(k-1)!(n-k)!}p^kq^{n-k}$$

$\ell=k-1, m=n-1$ とおき, $n!=n\times(n-1)!=n\times m!$, $\quad n-k=m-\ell$ となることに注意してから二項定理を用いると, 上の式の右辺は

$$\sum_{\ell=0}^{m}\frac{n\times m!}{\ell!(m-\ell)!}p^{\ell+1}q^{m-\ell}=np\sum_{\ell=0}^{m}\frac{m!}{\ell!(m-\ell)!}p^{\ell}q^{m-\ell}=np$$

となる.

$V=m_2-\mu^2$ よりまず 2 次元モーメント m_2 を計算する.

$$\begin{aligned}m_2=\sum_{k=0}^{n}k^2p(k)&=\sum_{k=0}^{n}k^2\frac{n!}{k!(n-k)!}p^kq^{n-k}\\&=\sum_{k=2}^{n}k(k-1)\frac{n!}{k!(n-k)!}p^kq^{n-k}+\sum_{k=1}^{n}k\frac{n!}{k!(n-k)!}p^kq^{n-k}\\&=\sum_{k=2}^{n}\frac{n!}{(k-2)!(n-k)!}p^kq^{n-k}+np\end{aligned}$$

$\ell=k-2, m=n-2$ とおき, $n!=n(n-1)m!, n-k=m-\ell$ となることに注意すると

$$=\sum_{\ell=0}^{m}n(n-1)p^2\cdot\frac{m!}{\ell!(m-\ell)!}p^{\ell}q^{m-\ell}+np$$

となり, ここで二項定理を用いると

$$=n(n-1)p^2+np$$

を得る. よって,

$$V = \sigma^2 = n(n-1)p^2 + np - (np)^2 = n(p - p^2) = npq$$

となる. 従って

二項分布の平均	:	$\mu = np$
二項分布の分散	:	$V = npq$
二項分布の標準偏差	:	$\sigma = \sqrt{npq}$

演習

(a) ある新薬の治癒率は 80 %であるとする. いま, 5 人の患者にこの薬を用いたとき, 治る人数を表す確率変数を X とする.

(1) $X = 4$ となる確率を求めよ.

(2) X の確率分布を求め, X の平均と標準偏差を計算せよ.

(b) 男（女）が生まれる確率を 1/2 (1/2) として, 5 人の子供をもつ家族で次の事象の起こる確率を求めよ.

(1) 5 人のうち少なくとも 4 人が男である.

(2) 男と女が少なくとも 1 人は含まれる.

(3) 5 人とも性別は同じである.

(c) ある自動車部品メーカーでは, 生産される部品の 1 箱（10 個の部品）には, たかだか 1 個の不良品しか含まれていないことを保証している. 過去の経験からこの工場の製造工程では 5 %の不良品を出すことが分かっているとき, 任意の 1 箱がこの保証を満たす確率を求めよ.

(iv) **幾何分布 (geometrc destribution)**

S の標本点が無限個つまり

$$S = \{0, 1, 2, 3, \ldots\}$$

の場合にも離散確率分布が定義できる. p を 0 以上 1 以下の実数とし

$$p(x) = p(1-p)^x, \quad x \in S$$

で定義される離散確率密度で定まる (離散) 確率分布を幾何分布という.

- 対応する試行

  表が出る確率が p, 裏が出る確率が $1-p$ であるコインを投げる試行を行う. 1 回目の試行で

  $$(※) \begin{cases} \text{表が出た場合そこで終了する.} \\ \text{裏が出た場合もう一度コインを投げる.} \end{cases}$$

  そして同様に (※) を表が出るまで繰り返し行う. このとき表がでるまでに裏がでた回数は幾何分布に従う. 等比級数の和の公式

  $$\sum_{x=0}^{\infty} r^x = \frac{1}{1-r}, \quad (\text{ただし } |r| < 1)$$

を用いると,

$$P(S) = \sum_{x=0}^{\infty} p(x) = p\sum_{x=0}^{\infty}(1-p)^x = p\frac{1}{1-(1-p)} = 1$$

が計算できる.

- 平均

  級数の公式

$$\sum_{x=1}^{\infty} xr^x = \frac{r}{(1-r)^2}, \quad (\text{ただし } |r|<1)$$

  を用いると,

$$\mu = \sum_{x=1}^{\infty} xp(x) = \sum_{x=1}^{\infty} xp(1-p)^x = p \times \frac{1-p}{p^2} = \frac{1-p}{p}.$$

- 分散

  級数の公式

$$\sum_{x=1}^{\infty} x^2 r^x = \frac{r(1+r)}{(1-r)^3}, \quad (\text{ただし } |r|<1)$$

  を用いると,

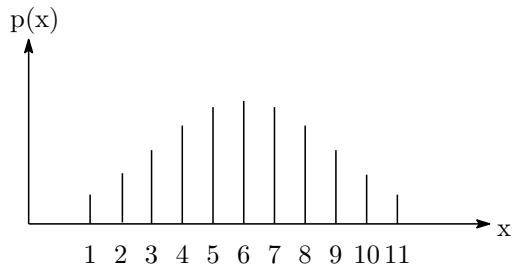
$$V = m_2 - \mu^2 = \sum_{x=1}^{\infty} x^2 p(1-p)^x - \left(\frac{1-p}{p}\right)^2 = p\cdot\frac{1}{p^3}(1-p)(2-p) - \frac{(1-p)^2}{p^2} = \frac{1-p}{p^2}$$

  従って,

幾何分布の平均	:	$\mu = \dfrac{1-p}{p}$
幾何分布の分散	:	$V = \dfrac{1-p}{p^2}$
幾何分布の標準偏差	:	$\sigma = \dfrac{\sqrt{1-p}}{p}$

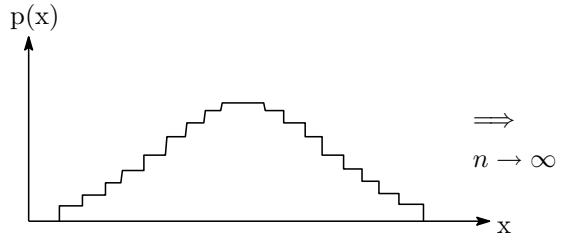
2.4.3 連続確率分布

離散分布（離散確率変数）との関係

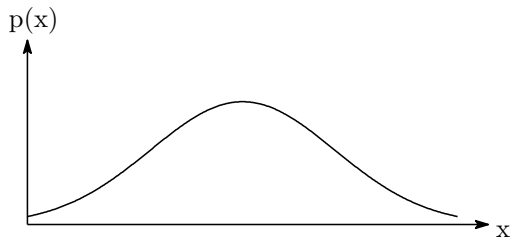


$$\sum_{i=1}^{n} p(x_i) = 1$$

x_1	x_2	x_3	$\cdots$	x_n
$p(x_1)$	$p(x_2)$	$p(x_3)$		$p(x_n)$



$\Longrightarrow$
$n \to \infty$



離散分布

連続分布

定義（連続確率密度）

p が $\mathbb{R} = (-\infty, \infty)$ 上の確率密度であるとは

(1) $p(x) \geq 0,\ x \in \mathbb{R}$

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

が成り立つ事である. A を $\mathbb{R}$ 上の事象としたとき

$$P(A) = \int_A p(x)dx$$

と定義すると P は確率となる.

連続確率分布の平均, 2 次モーメント, 分散, 標準偏差は次で定義される.

- **平均**

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

- **二次モーメント**

$$m_2 = \int_{-\infty}^{\infty} x^2 p(x)dx$$

- **分散**

$$V = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = m_2 - \mu^2$$

- **標準偏差**

$$\sigma = \sqrt{V}$$

2.4.4 連続確率分布の例

(i) 一様分布 (Uniform distribution)

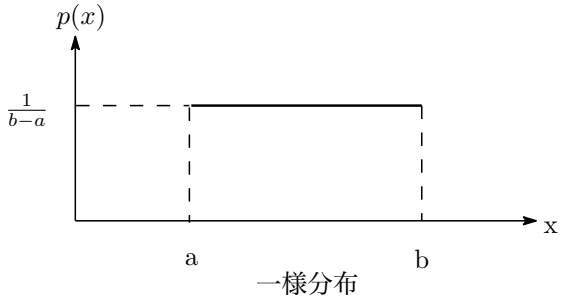
定義

$-\infty < a < b < \infty$ に対して $p(x)$ が

$$p(x) = \begin{cases} \dfrac{1}{b-a} & (a < x < b) \\ 0 & (\text{その他}) \end{cases}$$

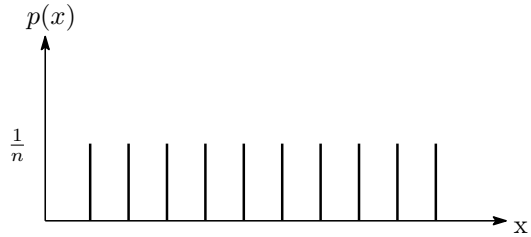
であるとき, $p(x)$ を確率密度とする分布を区間 (a, b) 上の一様分布という.

離散一様分布との比較



一様分布

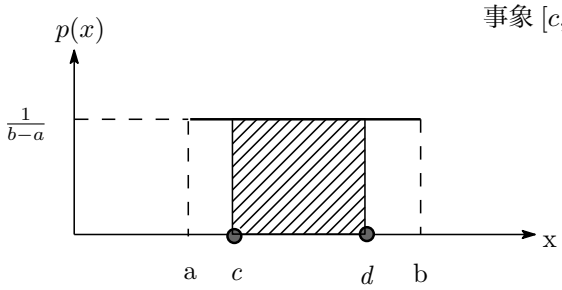
$$\begin{aligned} \int_{-\infty}^{\infty} p(x)dx &= \int_a^b \frac{1}{b-a}dx \\ &= \frac{1}{b-a}[x]_a^b \\ &= \frac{b-a}{b-a} \\ &= 1 \end{aligned}$$



離散一様分布

$$S = \{1, 2, \ldots, n\}$$
$$p(x) = \frac{1}{n}, \quad x \in S$$
$$\sum_{x=1}^{n} p(x) = 1$$

例) $a < c < d < b$ の時, 区間 $[c, d]$ の確率を計算してみる.



$$\begin{aligned} \text{事象 } [c,\, d] \text{ の確率} &= P(c < X < d) \\ &= \int_c^d p(x)dx \\ &= \frac{1}{b-a}\int_c^d dx \\ &= \frac{d-c}{b-a} \end{aligned}$$

一様分布の平均, 2次モーメント, 分散, 標準偏差の計算

- **平均**

$$\mu = \int_{-\infty}^{\infty} xp(x)dx = \frac{1}{b-a}\int_a^b xdx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{1}{2}\times\frac{b^2-a^2}{b-a} = \frac{a+b}{2}$$

- **2 次モーメント**

$$m_2 = \int_{-\infty}^{\infty} x^2 p(x)dx = \frac{1}{b-a}\int_a^b x^2 dx = \frac{1}{b-a}\times\frac{b^3-a^3}{3} = \frac{a^2+ab+b^2}{3}$$

- **分散**

$$V = m_2 - \mu^2 = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{4(a^2+ab+b^2)-3(a+b)^2}{12} = \frac{(a-b)^2}{12}$$

- **標準偏差**

$$\sigma = \sqrt{V} = \frac{a-b}{2\sqrt{3}}$$

(ii) **指数分布 (Exponential distribution)**

定義 指数分布: 確率密度関数 $p(x)$ が

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & (x \geq 0) \\ 0 & (x < 0) \end{cases}$$

である分布を (パラメータ $\lambda > 0$ の) 指数分布という.

対応するモデル　電球の寿命

時刻 t まで電球が切れていない時, 時刻 $t+h$ まで電球が切れない確率は t に無関係, つまり X を電球の寿命とすると

$$P(X \geq t+h | X \geq t) = P(X \geq h) = \int_h^{\infty} \lambda e^{-\lambda x}dx = e^{-\lambda h}$$

が成り立つ.

指数分布の平均, 2 次モーメント, 分散, 標準偏差を計算するためには次の部分積分の公式を用いる.

部分積分の公式

$$\int_a^b g(x)h(x)dx = [G(x)h(x)]_a^b - \int_a^b G(x)h'(x)dx$$

ここで $G(x)$ は, $g(x)$ の原始関数 ($G(x)$ の微分が $g(x)$) である.

- **平均**

$$\mu = \int_0^\infty p(x)xdx = \left[-xe^{-\lambda x}\right]_0^\infty + \int_0^\infty e^{-\lambda x}dx = \left[-\frac{1}{\lambda}e^{-\lambda x}\right]_0^\infty = \frac{1}{\lambda}$$

- **2 次モーメント**

$$m_2 = \int_0^\infty p(x)x^2dx = \lambda\int_0^\infty x^2e^{-\lambda x}dx = \left[-x^2e^{-\lambda x}\right]_0^\infty + 2\int_0^\infty xe^{-\lambda x}dx = \frac{2}{\lambda^2}$$

- **分散**

$$V = m_2 - \mu^2 = \frac{1}{\lambda^2}$$

- **標準偏差**

$$\sigma = \sqrt{V} = \frac{1}{\lambda}$$

(iii) **正規分布 (Normal distribution)**

定義

$\sigma > 0$, $-\infty < \mu < \infty$ とする. 確率密度関数 $p(x)$ が

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

である分布をパラメータ μ, σ の正規分布という. ($N(\mu, \sigma^2)$ と書く) 特に, $\mu = 0, \sigma = 1$ のときの正規分布を標準正規分布 (規準正規分布) という. このとき確率密度関数は,

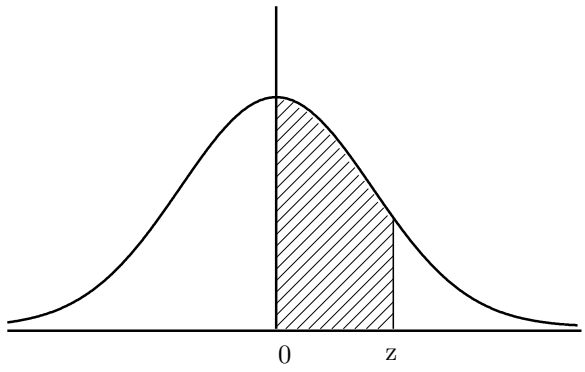
$$p(x) = p_{01}(x) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{x^2}{2}\right\}$$

となる. 区間 $[a, b]$ の確率

$$P_{\mu\sigma}([a, b]) = \int_a^b p_{\mu\sigma}(x)dx$$

は一般の a, b で積分は計算できない. そのため標準正規分布表を参考にする.

標準正規分布表は $F(z) = \displaystyle\int_0^{|z|} \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{x^2}{2}\right\}dx$ の値をまとめている表である.



例えば, $z = 1, 2, 3$ での $F(z)$ はそれぞれつぎの値である.

$$\begin{aligned} F(1) &= \int_0^1 \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx = 0.3413 \\ F(2) &= 0.4772 \\ F(3) &= 0.4987 \end{aligned}$$

$F(x)$ は次の性質を持つ.

$$F(z) = F(-z), \quad F(\infty) + F(-\infty) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx = 1, \quad F(\infty) = \frac{1}{2}$$

そして標準正規分布に対して区間 $[a, b]$ の確率は以下の性質を用いて表わすことができる.

$(0 \leq a < b)$ のとき

$$P_{01}([a, b]) = \int_a^b p_{01}(x)dx = \int_0^b p_{01}(x)dx - \int_0^a p_{01}(x)dx = F(b) - F(a)$$

$(a < b \leq 0)$ のとき

$$P_{01}([a, b]) = \int_a^b p_{01}(x)dx = \int_{-b}^{-a} p_{01}(x)dx = F(-a) - F(-b) = F(a) - F(b)$$

$(a \leq 0 \leq b)$ のとき

$$P_{01}([a, b]) = \int_a^b p_{01}(x)dx = F(a) + F(b)$$

例)

$$\begin{aligned} P_{01}([1, 2]) &= F(2) - F(1) = 0.4772 - 0.3413 = 0.1359 \\ P_{01}([-1, 2]) &= F(-1) + F(2) = 0.4772 + 0.3413 = 0.8185 \\ P_{01}([1, \infty)) &= F(\infty) - F(1) = 0.5 - 0.3413 = 0.1487 \end{aligned}$$

一般の μ と σ の場合は標準化（規準化）を用いて確率を計算する.

定理　（標準化（規準化））

X の分布がパラメータ μ, σ の正規分布 $N(\mu, \sigma^2)$ であるとき,
$Z = \dfrac{X - \mu}{\sigma}$ の分布は標準正規分布 $N(0, 1)$ である.

証明)

$z = \dfrac{x - \mu}{\sigma}$ として変数変換を行うと,

$$\begin{aligned} P(a \leq Z \leq b) &= P(a\sigma + \mu \leq X \leq b\sigma + \mu) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{a\sigma+\mu}^{b\sigma+\mu} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left\{-\frac{z^2}{2}\right\} dz. \end{aligned}$$

例)

$$\begin{aligned} &\int_\mu^{\mu+\sigma} p_{\mu\sigma}(x)dx = F(1) \\ &\int_\mu^{\mu+k\sigma} p_{\mu\sigma}(x)dx = F(k), \qquad k = 1, 2, \ldots, . \end{aligned}$$

平均, 分散, 標準偏差の計算は積分公式

$$(1) \quad \int_{-\infty}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} z dz = 0$$
$$(2) \quad \int_{-\infty}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} z^2 dz = \int_{-\infty}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} dz = \sqrt{2\pi}$$

を用いて示される. (1) は $z\exp\left\{-\frac{z^2}{2}\right\}$ が奇関数であり

$$\int_{-\infty}^{0} \exp\left\{-\frac{z^2}{2}\right\} z dz = -\int_{0}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} z dz$$

が成り立つことより得られ, (2) は $g(z) = z\exp\left\{-\frac{z^2}{2}\right\}, h(z) = z$ として部分積分の公式を用いると,

$$\int_{-\infty}^{\infty} z^2 \exp\left\{-\frac{z^2}{2}\right\} dz = \left[-\exp\left\{-\frac{z^2}{2}\right\} z\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} dz$$

となることより得られる.

- **平均**

  $z = \dfrac{x-\mu}{\sigma}$ で変数変換を行うと $(dz = \dfrac{1}{\sigma}dx, x = \mu + \sigma z)$

$$\begin{aligned}
\int_{-\infty}^{\infty} x p_{\mu\sigma}(x) dx &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma z) \exp\left\{-\frac{z^2}{2}\right\} dz \\
&= \mu \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{z^2}{2}\right\} dz + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z \left\{-\frac{z^2}{2}\right\} dz \\
&= \mu
\end{aligned}$$

  を得る. 最後の等号を導くために積分公式 (1) を用いた.

- **分散**

  $z = \dfrac{x-\mu}{\sigma}$ で変数変換を行うと,

$$\begin{aligned}
\int_{-\infty}^{\infty} (x-\mu)^2 p_{\mu\sigma}(x) dx &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-\mu)^2 \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sigma^2 z^2 \exp\left\{-\frac{z^2}{2}\right\} dz \\
&= \sigma^2
\end{aligned}$$

  を得る. 最後の等式を導くために積分公式 (2) を用いた. 従って,

$N(\mu, \sigma^2)$ の平均	:	μ
$N(\mu, \sigma^2)$ の分散	:	σ^2
$N(\mu, \sigma^2)$ の標準偏差	:	σ

演習

(i) 変数 Z が標準正規分布 $N(0,1)$ に従うとき, 標準正規分布表より以下の確率を求めよ.

(1) $Z > 1.26$　(2) $Z < -0.83$　(3) $|Z| < 0.97$　(4) $|Z| > 1.65$

(5) $P(Z > a) = 0.025$ となる a の値

(ii) 変数 X が正規分布 $N(69, 3^2)$ に従うとき, 標準化して以下の確率を求めよ.

(1) $X < 66$　(2) $65 < X < 71$　(3) $|X - 69| > 3$

(iii) 英語の試験の得点分布は平均 130 点, 標準偏差 20 点の正規分布にほぼ近い形をしていた.

(1) 100 点以上を合格とするとき, この試験で合格になる学生の割合はいくらか?

(2) 上位 2.5 % に奨学金を与えたい. 何点以上とすべきか?

2.5 大数の法則と中心極限定理

2.5.1 無作為抽出

無作為抽出とは各々の個体が標本に選ばれる確率が等しい抽出法である. これを正確に述べると次の様になる. 母集団の大きさを N, 標本の大きさを n とする.

- **$n = 1$ のとき**

  母集団から 1 つの標本をとるとき, どの個体も標本に選ばれる確率が $\frac{1}{N}$ になる抽出法.

- **$n = 2$ のとき**

  2 つの個体の組 $\{x_1, x_2\}$ が選ばれる確率が $\frac{2}{N(N-1)}$ になる抽出法.

- **一般の n のとき**

  n 個の個体の組 $\{x_1, x_2, \cdots, x_n\}$ が選ばれる確率が

  $$\frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!}$$

  になる抽出法.

＜注意＞ $N \gg n$ (n に比べて N が十分大きいとき)

$$\frac{n!(N-n)!}{N!} \fallingdotseq n! \times (\frac{1}{N})^n$$

となる. つまり n 個の個体の組を一度に無作為抽出した場合と (試行 3, 試行 4, 試行 5 参照), 復元抽出法を n 回行った場合とがほぼ同じ結果となる.

2.5.2 大数の法則

まず乱数表 (0 から 99 までの数が無作為に並べられている表) を用いて次の実験を行う.
各行の偶数の数を数えて, 度数, 相対度数を計算する.

	度数	相対度数	累積相対度数
1 行目の偶数の数	$f_1 = 9$	$\frac{9}{25} = 0.36$	$\frac{9}{25} = 0.36$
2 行目の偶数の数	$f_2 = 12$	$\frac{12}{25} = 0.48$	$\frac{21}{50} = 0.42$
3 行目の偶数の数	$f_3 = 16$	$\frac{16}{25} = 0.64$	$\frac{37}{75} = 0.49$
4 行目の偶数の数	$f_4 = 15$	$\frac{15}{25} = 0.60$	$\frac{52}{100} = 0.52$
5 行目の偶数の数	$f_5 = 12$	$\frac{12}{25} = 0.48$	$\frac{64}{125} = 0.512$

従って相対度数は, 標本の数を大きくしていくと本来の確率である $\frac{1}{2}$ に近づくことが分かる. 一般に次の定理が成り立つ.

定理 (大数の法則) n を大きくしていくと

(1) 標本分布は母集団分布に近づく.

(2) 相対度数は確率に近づく.

(3) 標本平均は母集団平均に近づく.

(4) 標本標準偏差は母集団標準偏差に近づく.

2.5.3 中心極限定理

- $X_1, X_2, \ldots, X_n$ の標本平均

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

は大数の法則により $n \to \infty$ で平均 EX_1 に収束する. n を固定すると $\overline{X}$ の分布はもとの $X_i, i = 1, 2, \ldots, n$ の分布に依存するが, n が十分大きいときどのように EX_1 に収束するかが分かる. つまり平均からのばらつきが計算できる.

定理 (中心極限定理)

X が $\mu_X = \mu, \sigma_X = \sigma$ である分布に従うとき, 大きさ n の標本平均 $\overline{X}$ は n が大きくなると $\mu_{\overline{X}} = \mu, \sigma_{\overline{X}} = \sigma/\sqrt{n}$ の正規分布に近づく. (n が 25 以上くらいであれば十分よい近似となり適用できる.)

(例)

米国人男性の身長 X は $\mu_X = 68$(インチ), $\sigma_X = 3$(インチ) である. 大きさ $n = 25$ の標本平均 $\overline{X}$ と μ_X の差が 1 インチ以内である確率 $P(|\overline{X} - \mu_X| \leq 1) = P(67 \leq \overline{X} \leq 69)$ を中心極限定理を用いて求めよ.

(証明)

中心極限定理を用いると, $\overline{X}$ は平均 $\mu_{\overline{X}} = 68$, 標準偏差 $\sigma_{\overline{X}} = \frac{3}{\sqrt{25}} = 0.6$ の正規分布で近似できる. 従って規準化を行うと

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - 68}{0.6}$$

は標準正規分布に従う. ゆえに,

$$P(|\overline{X}-\mu|\le 1)=P(67\le\overline{X}\le 69)=P\left(\frac{-1}{0.6}\le\frac{\overline{X}-68}{0.6}\le\frac{1}{0.6}\right)$$
$$\fallingdotseq P(-1.67\le Z\le 1.67)\fallingdotseq 0.4525\times 2=0.905$$

従って, 標本平均が 67 インチ以上 69 インチ以下である確率は 90.5% である.

演習

前の例で母集団平均 μ_X は未知であるとする. すなわち X は $\mu_X=\mu$(未知), $\sigma_X=3$ の任意の分布に従うとする. 大きさ $n=25$ の標本平均 $\overline{X}$ の値と μ の差が 1 インチ以内である確率を求めよ.

2.5.4 2 項分布の正規近似

X をパラメータ n,p の二項分布に従う確率変数とする.

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}=\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}\qquad(k=0,1,2,\dots,n)$$

n が大きいとき, 2 項分布の確率の計算は困難である. そこで正規分布で近似を行う. 成功の割合を $\hat{p}$ とおく. つまり

$$\hat{p}=\frac{X}{n}.$$

$X,\hat{p}$ の平均, 分散, 標準偏差は

	X	$\hat{p}$
平均	np	p
分散	$np(1-p)$	$\dfrac{p(1-p)}{n}$
標準偏差	$\sqrt{np(1-p)}$	$\sqrt{\dfrac{p(1-p)}{n}}$

となる. 大数の法則より

$$\hat{p}\to p,\quad n\to\infty$$

が成り立ち, さらに中心極限定理より n が大きいとき ($n\ge 200$), $\hat{p}-p$ が正規分布 $N\left(0,\dfrac{p(1-p)}{n}\right)$ で近似できる. つまり任意の $a,b\in\mathbb{R}$ $(a<b)$ に対して

$$P(a<\hat{p}-p<b)\fallingdotseq P\left(a<\sqrt{\frac{p(1-p)}{n}}Z<b\right)=P\left(\frac{a\sqrt{n}}{\sqrt{p(1-p)}}<Z<\frac{b\sqrt{n}}{\sqrt{p(1-p)}}\right)$$

となる. ここで Z は標準正規分布 $N(0,1)$ に従う確率変数である.

演習

(i) ある都市のドライバーの 20 % は 1 年に少なくとも 1 回の事故を起こすとする. この都市のある自動車損害保険会社の契約者 200 人中, 事故を起こす人の割合を $\widehat{p}$ とする.

(1) $\widehat{p}$ の平均と標準偏差を求めよ.

(2) $\widehat{p}>0.25$ となる確率を正規分布で近似して求めよ.

3　推測統計

母集団について完全に知ることはしばしば困難である。たとえば
(i)　日本人全体が調査対象であるといったように母集団が非常に多くの要素（無限大の場合も含む）からなる場合.
(ii)　母集団は, あまり多くの要素は含まないが, 個々の調査が高価であるため予算上の制約から全数の調査が不可能である場合.
(iii)　未来の経済成長率のように将来起こるため, 現在の測定が不可能な要素を含む場合.

などがある. このようなとき, (a) 母集団からその一部を選び出し, (b) それを分析し, (c) 母集団について推測する, ということが行われる. これを記述統計にたいして **推測統計** (statistical inference) と呼ぶ.

推測統計において知りたいものは, もとの母集団についてである. 母集団はその分布を持っているので, 目的としてその分布を知ればよい。これを **母集団分布** (population ditribution) という. 統計学では, どちらかといえば無限母集団 ($N=\infty$) を考えることが多いので, 母集団分布は, 確率分布（密度）$p(x)$ と考える. そして「標本 $X_1, X_2, \ldots, X_n$ は同一の母集団分布 $p(x)$ に従う n 個の独立な確率変数である.」ということがいえる. このとき, $p(x)$ は問題に応じて連続型でも離散型でもよい.

母集団分布がある知られた確率分布であることが, 理論的・経験的に分かっている場合、いくつかの定数が分かれば母集団分布についてすべて知ることができる. この場合は母集団分布を決定する定数 (この定数を母数 (parameter) と呼ぶ) を求めることが目的である. 母集団分布 $p(x)$ を特定する代表的な母数は, その平均

$$\mu = \int_{-\infty}^{\infty} xp(x)dx, \text{ あるいは } \mu = \sum_x xp(x),$$

である. これを **母平均** (population mean) という. 同じようにして **母分散** (population variance) も定まる. $p(x)$ の値をすべて知らなくても母平均 μ と 母分散 σ^2 を知れば多くのことを知ることができる. 特に正規分布の場合は, 母平均 μ と 母分散 σ^2 が分かれば $p(x)$ の値すべてが分かることになる.

各々の分布に対する母数
(i)　二項分布では, 試行回数 n と 1 回の試行における成功の確率 p によって分布が決まる. したがって, 組 (n,p) が母数となる.
(ii)　ポアソン分布は, 平均 $\lambda > 0$ によって分布が定まる. (標準偏差も λ であることに注意.) したがって, λ が母数となる.
(iii)　一様分布では, 範囲 $-\infty < a < b < \infty$ によって分布が定まる. したがって, 組 (a,b) が母数となる.
(iv)　正規分布では, 平均 μ と 標準偏差 σ によって分布が定まる. したがって, 組 (μ,σ) が母数となる.

3.1　推定量

未知の母数の値を観測値に基づいて求める問題を統計的推定 (statistical estimation) と呼ぶ. 一般に推定のために用いられる統計量を**推定量**という. そして推定量に標本観測値を代入して得られる値を**推定値**という.

推定量の例
(i)　標本平均 $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ は, 母集団平均 $\mu = EX$ の推定量である.
(ii)　母標準偏差 $\sigma = \sqrt{E[(X-\mu)^2]}$ の推定量として次で定義された二通りの標本標準偏差がある.

$$S_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}, \quad \widehat{S}_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

定義 (不偏推定量)
推定量の期待値が推定しようとしている母数の値と一致しているときその推定量を**不偏推定量**という.

注意　不偏性のない推定量を用いることは, クセのある推定法を用いることであり, 望ましくない.

定理

(1)　標本平均 $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ は μ の不偏推定量である.

(2)　標本分散 $S_X^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ は σ^2 の不偏推定量である.

証明　まず

$$E[\overline{X}] = \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu$$

より (1) が示される. 次に $E[S_X^2] = \sigma^2$ を示す. 期待値の線形性より

$$E[S_X^2] = \frac{1}{n-1}\sum_{i=1}^{n} E[(X_i - \overline{X})^2] = \frac{1}{n-1}\left\{\sum_{i=1}^{n} E[X_i^2] - 2\sum_{i=1}^{n} E\left[X_i\overline{X}^2\right] + nE\left[\overline{X}^2\right]\right\}.$$

となる. $X_i, i = 1, 2, \ldots, n$ が同分布であることから $\sum_{i=1}^{n} E[X_i^2] = nE[X_1^2]$ である. また $X_i, i = 1, 2, \ldots, n$ が独立同分布であるので, $i \neq j$ のとき $E[X_iX_j] = E[X_i]E[X_j] = E[X_1]^2$ が成り立つことを用いると

$$\begin{aligned}\sum_{i=1}^{n} E[X_i\overline{X}] &= \sum_{i=1}^{n} E\left[X_i\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} E[X_iX_j] \\ &= \frac{1}{n}\Big\{\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} E[X_iX_j] + \sum_{i=1}^{n} E[X_i^2]\Big\} = (n-1)E[X_1]^2 + E[X_1^2].\end{aligned}$$

と

$$\sum_{i=1}^{n} E[\overline{X}^2] = E\left[\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)^2\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E[X_iX_j] = \frac{1}{n}\Big\{(n-1)E[X_1]^2 + E[X_1^2]\Big\}.$$

が導かれる. 以上をまとめると

$$\begin{aligned}E[S_X^2] &= \frac{1}{n-1}\left\{\sum_{i=1}^{n} E[X_i^2] - 2\sum_{i=1}^{n} E\left[X_i\overline{X}^2\right] + nE\left[\overline{X}^2\right]\right\} \\ &= \frac{1}{n-1}\Big\{nE[X_1^2] - 2(n-1)E[X_1]^2 - 2E[X_1^2] + (n-1)E[X_1]^2 + E[X_1^2]\Big\} \\ &= E[X_1^2] - E[X_1]^2 = \sigma^2.\end{aligned}$$

をえる.

一致性　一致性は, 標本の大きさが n が大きくなるに従い推定量 $\widehat{\theta}_n$ が真の母数の値 θ に近づく性質である. 数学的には, どのような正の実数 ε にたいしても

$$n \to \infty \text{ のとき} \quad P(|\widehat{\theta}_n - \theta| > \varepsilon) \to 0$$

となることである. この条件を満足する場合, $\widehat{\theta}_n$ は **一致推定量** (consistent estimator) であるという.

3.2 推定

3.2.1 点推定と区間推定

- **点推定**： 母集団の未知の母数 θ を推定する場合, ある 1 つの値（推定量）$\widehat{\theta}$ で推定する方法を**点推定** (point estimation) と呼ぶ.
  - 平均 μ を推定するとき, $\widehat{\theta}$ は
  $$\text{標本平均} : \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
  - 標準偏差 σ を推定するとき, $\widehat{\theta}$ は
  $$\text{標本標準偏差} : S_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

  $\widehat{\theta}$ は $X_1, X_2, \ldots, X_n$ の関数である確率変数である. したがって, 現実の $\widehat{\theta}$ の値（推定値）は θ に一致せず, 実際の推定にはある程度の誤差を伴う. よって誤差の評価を行わなければならないが, これには確率的な取り扱いが必要である.

- **区間推定**：**区間推定** (interval estimation) は, パラメータの真の値が入る確率がある値 $1-\alpha$ 以上と保証される区間 $[L, U]$ を求めるもの, つまり
$$P(L \leqq \mu \leqq U) \geqq 1 - \alpha$$
なる L, U を求めるものであり, 最初からある程度の誤りがあることを認めた推定法といえる. L, U は $X_1, X_2, \ldots, X_n$ の関数, つまり統計量, 左辺の確率が $1-\alpha$ 以上になるように標本分布から求められる.
  - $1-\alpha$ を**信頼係数** (confidence coefficient) といい.
  - L, U をそれぞれ, **下側信頼限界** (lower confidence limit), **上側信頼限界** (upper confidence limit) といい,
  - 区間 $[L, U]$ を $100(1-\alpha)$ パーセント**信頼区間** (confidence interval) という.

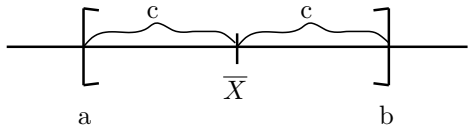
  $(1-\alpha)$ としては目的に応じた適当な値を選ぶが, $0.99, 0.95, 0.90$ に設定されることが多い.

3.2.2 任意の分布での母集団平均 μ の推定

(1) 母集団分布は任意で標準偏差 σ が**既知の場合**

標本の数 $n \geqq 25$ の時, 中心極限定理より標本平均 $\overline{X}$ は平均 $\mu_{\overline{X}} = \mu$（未知), 標準偏差 $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$ の正規分布で近似できる.

目的　母数が含まれる確率が 95% である区間 $[a, b]$ を求める.



$P(\overline{X} - c \leqq \mu \leqq \overline{X} + c) \fallingdotseq 0.95$ なる c を求める.

$$\text{左辺} = P(|\mu - \overline{X}| \leqq c) = P\left(\left|\frac{\mu - \overline{X}}{\sigma_{\overline{X}}}\right| \leqq \frac{c}{\sigma_{\overline{X}}}\right) \fallingdotseq P\left(|Z| \leqq \frac{c}{\sigma_{\overline{X}}}\right)$$

↑ 中心極限定理の規準化

$$Z = \frac{\mu - \overline{X}}{\sigma_{\overline{X}}}, \quad \left(\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\right)$$

従って $P(|Z| \leqq x_{0.95}) \fallingdotseq 0.95$ となる $x_{0.95}$ が標準正規分布表より 1.96 であることが分かるので

$$c = \sigma_{\overline{X}} x_{0.95} = \frac{\sigma}{\sqrt{n}} x_{0.95}$$

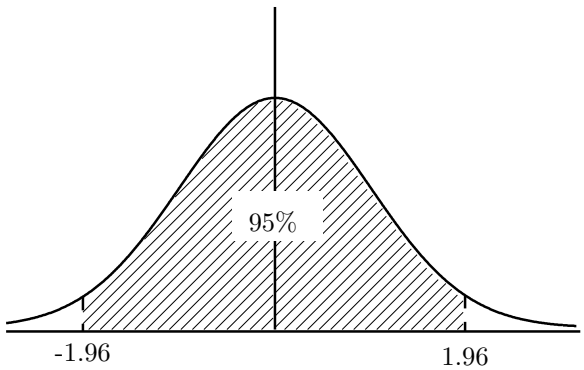
となることが導かれる. したがって μ が 95% で含まれる区間 (95% 信頼区間) は

$$\left[\overline{X} - 1.96 \times \frac{\sigma}{\sqrt{n}},\ \overline{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right]$$

である. もし 90% 信頼区間を求めるときには $X_{0.90} = 1.645$ つまり

$$\left[\overline{X} - 1.645 \times \frac{\sigma}{\sqrt{n}},\ \overline{X} + 1.645 \times \frac{\sigma}{\sqrt{n}}\right]$$

とすればよい.



(2) 母集団分布が任意で標準偏差 σ が**未知の場合**

- $\sigma_{\overline{X}}$ の代わりに標本標準偏差 $S_{\overline{X}}$ を用いる. $(\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}, S_{\overline{X}} = \frac{S_X}{\sqrt{n}})$
- **近似定理**

  大きさ $n(n \geqq 100)$ の標本をとると, 標本平均 $\overline{X}$ は平均 $\mu_{\overline{X}} = \mu$, 標準偏差 $S_{\overline{X}}$ の正規分布で近似できる.

  (注) σ の点推定は S_X である. $n \geqq 100$ の時, その誤差は無視できる.

上の近似定理より次の公式が導かれる.

μ の (近似的な)95% 信頼区間は

$$[\overline{X} - 1.96 \times \frac{S_X}{\sqrt{n}}, \overline{X} + 1.96 \times \frac{S_X}{\sqrt{n}}]$$

である. そして, μ の (近似的な)90% 信頼区間は

$$[\overline{X} - 1.645 \times \frac{S_X}{\sqrt{n}}, \overline{X} + 1.645 \times \frac{S_X}{\sqrt{n}}]$$

である.

例題 I ある学校で 100 人の生徒の無作為標本が選ばれ, これらの生徒に知能テストが行われた. テストの結果 100 人の生徒の知能指数が決まり, $\overline{x} = 112, s_x = 11.0$ が得られた. これらの標本値を基にして, この学校の生徒全体の知能指数に対する 95% 信頼区間を求めよ.

【解】　上の公式に標本値を代入すると

$$\begin{aligned} \overline{X} - 1.96 \times \frac{S_X}{\sqrt{n}} &\leqq \mu \leqq \overline{X} + 1.96 \times \frac{S_X}{\sqrt{n}} \\ 112 - 1.96 \times \frac{11.0}{\sqrt{100}} &\leqq \mu \leqq 112 + 1.96 \times \frac{11.0}{\sqrt{100}} \\ 109.844 &\leqq \mu \leqq 114.146 \\ 109.8 &\leqq \mu \leqq 114.1 \end{aligned}$$

をえる.

3.2.3 母集団の成功の確率 p の推定

1 回の成功の確率が p の試行を n 回繰り返す. 標本からの成功の割合を $\widehat{p}$ と書く. つまり $\widehat{p} = \dfrac{1}{n} \times$成功の回数である. 成功の回数 Σ_n は二項分布 $B(n,p)$ に従うことがわかる. したがって $\mu_{\Sigma_n} = E[\Sigma_n] = np$, $\sigma_{\Sigma_n} = \sqrt{E[\Sigma_n^2] - \mu_{\Sigma_n}^2} = \sqrt{np(1-p)}$ であり,

$$\widehat{p} = \frac{\Sigma_n}{n} \quad \text{の平均は} \quad \mu_{\widehat{p}} = \frac{\mu_{\Sigma_n}}{n} = p, \quad \text{標準偏差は} \quad \sigma_{\widehat{p}} = \frac{\sigma_{\Sigma_n}}{n} = \sqrt{\frac{p(1-p)}{n}}$$

となる.

- **近似定理**

  大きさ $n(n \geqq 100)$ の標本をとり, 成功の割合を $\widehat{p}$ とおく. $\widehat{p}$ の分布は平均 p, 標準偏差 $\sqrt{\dfrac{\widehat{p}(1-\widehat{p})}{n}}$ の正規分布で近似される.

母集団の成功の確率 p に対する 95% 信頼区間は

$$\widehat{p} - 1.96 \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leqq p \leqq \widehat{p} + 1.96 \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

となり, 90% 信頼区間は

$$\widehat{p} - 1.645 \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leqq p \leqq \widehat{p} + 1.645 \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

例題 II ある都市で 1 日に少なくとも 1 箱のタバコを吸う成人男性の割合 p を推定したいと考えた. 大きさ 300 の無作為標本を採って調べた結果, この様な喫煙者が 36 人いた. p の 95% 信頼区間を求めよ.

【解】　$\widehat{p} = \dfrac{\Sigma_n}{n} = \dfrac{36}{300} = 0.12$ を公式に代入すると

$$\begin{aligned} 0.12 - 1.96\sqrt{\frac{0.12 \cdot 0.88}{300}} &\leqq p \leqq 0.12 + 1.96\sqrt{\frac{0.12 \cdot 0.88}{300}} \\ 0.083 &\leqq p \leqq 0.157 \end{aligned}$$

をえる.

3.2.4 標本の大きさの問題

- 平均 μ の推定誤差 $|\mu - \overline{X}|$ がある値 ε を超えない確率を 95% 以上にしたい. 標本の大きさ n をいくつ以上にすべきか？

  【解法】

  $$\overline{X} - 1.96 \cdot \frac{S_X}{\sqrt{n}} \leqq \mu \leqq \overline{X} + 1.96 \cdot \frac{S_X}{\sqrt{n}}$$

  つまり $|\mu - \overline{X}| \leqq 1.96\dfrac{S_X}{\sqrt{n}}$ が 95% で起こる. 従って

  $$\begin{aligned} 1.96 \times \frac{S_X}{\sqrt{n}} &\leqq \varepsilon \\ &\Downarrow \\ \left(\frac{1.96 S_X}{\varepsilon}\right)^2 &\leqq n \end{aligned}$$

  とすればよい.

- 成功の確率 p の推定誤差 $|p - \widehat{p}|$ がある値 ε を超えない確率を 95% 以上にしたい. 標本の数 n はいくつ以上にすべきか？

  【解法】

  $$\widehat{p} - 1.96 \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leqq p \leqq \widehat{p} + 1.96 \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

  つまり $|p - \widehat{p}| \leqq 1.96 \times \sqrt{\dfrac{\widehat{p}(1-\widehat{p})}{n}}$ が 95% で起こる. 従って

  $$\begin{aligned} 1.96 \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} &\leqq \varepsilon \\ &\Downarrow \\ \left(\frac{1.96}{\varepsilon}\right)^2 \widehat{p}(1-\widehat{p}) &\leqq n \end{aligned}$$

  とすればよい.

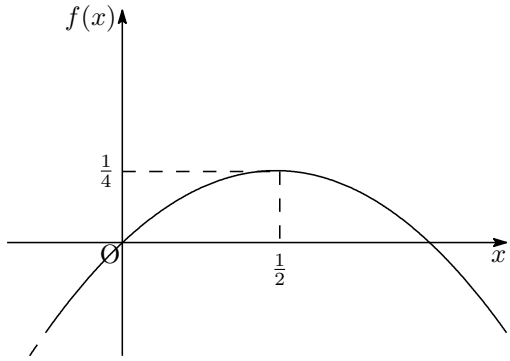
- $\widehat{p}$ について何の情報も無い時, $\widehat{p}$ のかわりに $\frac{1}{2}$ を使い,

  $$\frac{1}{4}\left(\frac{1.96}{\varepsilon}\right)^2 \leqq n$$

  とすればよい. 理由は, $f(x) = x(1-x)$ の $0 \leqq x \leqq 1$ での振る舞いを調べると

  $$f(x) = x(1-x) = -x^2 + x = -(x - \tfrac{1}{2})^2 + \tfrac{1}{4}$$

  であることから, f は $x = \dfrac{1}{2}$ の時最大で最大値 $\dfrac{1}{4}$ をとるからである.



注意

(1) 90% 以上にしたいときは, 1.96 を 1.645 に変更すればよい.

(2) 誤差を $\dfrac{1}{10}$ にしたい場合は標本の数は 100 倍にする必要がある.

3.2.5 スチューデントの t 分布

(σ が未知, 正規分布, 任意の n)

- μ の推定で n が小さく ($n < 25$), σ が未知の時は大標本法（中心極限定理を用いたもの) を用いることができない. 一般にはこの場合は困難. もし”X が正規分布に従う”という条件があれば次の 小標本法 を用いる事ができる.

定理 母集団分布が正規分布で, 標準偏差 σ が未知のとき, 大きさ n の標本をとると

$$t = \frac{\overline{X} - \mu}{S_X}\sqrt{n}$$

は自由度 $\nu = n - 1$ のスチューデント分布 t に従う.

小標本法による μ の 95% 信頼区間は

$$\overline{X} - t_0\frac{S_X}{\sqrt{n}} \leq \mu \leq \overline{X} + t_0\frac{S_X}{\sqrt{n}}$$

ただし t_0 はスチューデントの t 分布表における $\nu = n - 1$, $p = 0.025$ の値.

例題 III ある製パン工場で作るパンの重さは正規分布に従うことがわかっている. 検査員がパンの重さを検査するために, 15 個の無作為標本をとった. これら 15 個のパンの重さの平均と標準偏差は 15.8 オンスと 0.3 オンスであった. μ の 95% 信頼区間を求めよ.

【解法】　$t_{0.025}(14)$ の値を求める.

$$\left.\begin{array}{rcl} \nu = n - 1 = 15 - 1 & = & 14 \\ p = (1 - 0.95)/2 & = & 0.025 \end{array}\right\} \longrightarrow t_{0.025}(14) = 2.145$$

従って $s_x = 0.3$, $\overline{x} = 15.8$ を代入すると

$$\begin{array}{rcl} 15.8 - 2.145 \times \dfrac{0.3}{\sqrt{15}} & \leq \mu \leq & 15.8 + 2.145 \times \dfrac{0.3}{\sqrt{15}} \\ \longrightarrow \quad 15.63 & \leq \mu \leq & 15.97 \end{array}$$

となる.

3.2.6 推定の演習問題

(i) ある地域の住民 400 人の無作為標本のうち 280 人は虫歯予防のため水道の水に少量のフッ素化合物を入れることを希望するという結果が出た． このデータからフッ素化合物の混合を希望する人の割合の 95% 信頼区間を求めよ．

(ii) 新聞社は消費税を支持する有権者の割合 p を推定しようと考えている．0.90 の確率で推定値 $\widehat{p}$ を真の割合 p との差が 0.03 以内になるには標本を何人以上とらなければならないか？もし予備知識として $\widehat{p} \fallingdotseq 0.2$ を持っていればどれだけの標本でいいのか？

(iii) ある保険会社は長い間の経験から傷害保険加入者の 30% が 3 年間に少なくとも 1 回自動車事故を起こすことを知っている．この会社は, ある市の職員全員がこの傷害保険に加入する事を予期して, その場合の保険料率を決めたいと考えた．そこで, この市職員の中から 100 人の標本をとり, 過去 3 年間に少なくとも 1 回事故を起こした者を調べたら, 25 人がそうであった．市職員は全保険加入者を代表すると仮定して, 次の問題を解け．

(a) この推定値の確率的な精度はいくらか．推定値の誤差 ε を超えない確率が 95 パーセントとなるような ε の値を求めなさい．実際の精度はいくらか．そして, 実際の精度が標本割合は理論から期待される結果に矛盾するかどうかを確かめなさい．

(b) 0.95 の確率で, 推定値の誤差を 0.03 以内にとどめるためには, 標本をさらに何人追加しなければならないか．p は未知であるとして解け．

(c) (b) の問題を, p の標本推定値を使う代わりに, 控えめな値 $p = 1/2$ を用いて解け．

(d) p の値は未知であるとして, p に対する 90% 信頼区間を求めよ．この区間は実際に p を含むか．

(iv) 1972 年秋発刊の医学雑誌 *Canadian Medical Association Journal* に, "ビタミン C の大量投与が風邪に及ぼす効果について" という題の論文が掲載されている．この論文は, ビタミン C には風邪をはやく治す効果が若干あるという結論を下している．風邪をひいて, ビタミン C の大量投与を受けた 407 人についての実験では, 風邪が治るまでに要した日数の平均と標準偏差はそれぞれ 5.25 日と 6.0 日であった．標準偏差がこんなに大きいのは, おそらく, 風邪が治るまでに長い日数を要する人が必ず何人かいて, そのため風邪の回復に要する日数の分布は右側に長いすそを持つためである．このデータを用いて, 風邪が治るまでの日数の分布の平均値に対する 90% 信頼区間を求めよ．

(v) ある型の自動車の走行距離を推定するため, その型の車 30 台を標本に選び, 1 台ずつテストを行なった．30 台の走行距離の平均と標準偏差がそれぞれ 19.6 マイルと 0.7 マイルになったとして, この型の車の平均走行距離に対する 90% 信頼区間を求めよ．

(vi) ある部品の生産者は, その製品には約 3% の不良品があると思っている．いま真の不良率を 0.97 の確率で 0.5% まで正確に推定したいとすれば, どれだけの標本をとらねばならないか．

(vii) 60 匹の実験動物を 2 週間, ある種の餌を与えて飼育した．そのときの体重増のデータから, $x = 42$ オンス, $s = 4$ オンスが得られた．

(a) 母集団平均の推定値として 42 オンスはどのくらい正確であるか．推定値の誤差 ε を超えない確率が 95 パーセントとなるような ε の値を求めなさい．

(b) $\bar{x}$ と μ の違いを 0.95 の確率で 1/2 オンス以下にするためには, どれだけの標本が必要か．

(c) μ に対する 95% 信頼区間を求めよ．ここでは大標本法を用いよ．

(viii) 学生自治会は新しい学則を支持する学生の割合を推定しようとしている．300 人の学生からなる無作為標本を選ぶことが提案された．これらの学生に対する質問の結果 $\hat{p} = 0.60$ を得た．

(a) この値は真の割合の推定値としてどのくらい正確といえるか. 推定値の誤差 ε を超えない確率が 95 パーセントとなるような ε の値を求めなさい .

(b) p の推定値を 0.04 以内の誤差で求めたいとすれば, 自治会は何人の標本学生をとればよいか. 0.95 の確率であれば十分であるとして, $\hat{p} = 0.60$ を用いて解け.

(c) 自治会が p の推定値として $\hat{p} = 0.60$ という予備知識を持っていなかったとすれば, どれだけの標本をとらねばならないか.

(ix) x が正規分布に従うとき, $x = 20, s = 4, n = 12$ が与えられたとして, スチューデントの t 分布を用いて

(a) μ に対する 95% 信頼区間を求めよ.

(b) μ に対する 99% 信頼区間を求めよ.

(x) ある銘柄のタバコ 20 本からなる標本について, そのニコチン含有量を調べ, $\bar{x} = 22(mg), s = 4(mg)$ を得た.

(a) スチューデントの t 分布を用いて, μ に対する 95% 信頼区間を求めよ.

(b) 大標本法によって解き, 大標本法と小標本法の結果を比較せよ.

(xi) ある種のガソリン添加物が自動車の走行距離を延ばすかどうかを調べるためのテストを行った. 25 台の車にそれぞれガソリン 5 ガロンを給油し, ガソリンが無くなるまで車を走らせた. 実験完了後, 各車ごとに 1 ガロンあたりの走行距離を計算した. こうして求めた 25 台の車の 1 ガロンあたり走行距離の平均と標準偏差はそれぞれ $\bar{x} = 18.5$ マイル, $s = 2.2$ マイルであった. 添加物を加えないで, 同じ種類の自動車を使い, 長期間にわたり行なってきたこれまでのテストの経験では, 1 ガロンあたり走行距離の平均と標準偏差は, 約 $\mu = 18.0$ マイル, $\sigma = 2.0$ マイルであった. この添加物は車の走行距離に影響しないと仮定して, 次の問題を解け.

(a) μ の推定値 $\bar{x}$ の確率的な精度を求めよ. また, $\bar{x}$ の実際の精度はどうなるか. 標本値は, 理論から期待される結果に矛盾しないか.

(b) 推定値の誤差が 1 ガロンあたり 1/2 マイルを超えない確率を 0.95 とするには, 何台の車でテストを行わねばならないか.

(c) μ に対する 95% 信頼区間を求めよ. 求めた区間は実際に μ を含んでいるか.

(d) 添加物は平均, 分散のいずれにも影響しないという仮定をはずした上で, スチューデントの t 変数を使い, μ に対する 95% 信頼区間を求めよ.

3.3 仮説検定

「仮説検定」は, 統計的仮説の「有意性」の検定である. 仮説の下でわれわれが期待するものと観測した結果との違いが, 単に「偶然」によって起こったものか否かという見地から確率の基準で評価する. 仮説検定は推定とならんで統計的推測の理論の双璧であるとともに, 他方で, 統計的判断の論理学, 科学方法論という意味をもつ.

3.3.1 平均の検定

例題 1 (片側検定) 電球 A, B の平均寿命について

今まで使っている電球 A は平均 1180h, 標準偏差 90h の寿命を持つことがわかっている. セールスマンがやってきて「電球 A と同じ品質ですよ」と言って価格が安い電球 B を売りにきた. もし A と B が同じ品質であれば B を使いたい. B から大きさ $n = 100$ の標本平均 $\bar{x} = 1140h$ を得た. ただし B の電球の平均寿命の標準偏差は 90h とする.

【仮説 (Hypothesis)】

電球 A と電球 B は同じ寿命を持つ $(\mu_A = \mu_B)$

【対立仮説】

(i) 電球 A と電球 B の寿命は異なる. $(\mu_A \neq \mu_B)$　両側検定
(ii)　電球 A のほうが電球 B より寿命が長い. $(\mu_A > \mu_B)$ ⎫
(iii)　電球 B のほうが電球 A より寿命が長い. $(\mu_B > \mu_A)$ ⎭ 片側検定

今回の問題では事前の知識（セールスマンの話）より対立仮説として (ii) を用いる. 事前の知識が無い場合, 両側検定を用いる. なにかの事前の知識がある場合どちらかの片側検定を行う. したがって

$$\text{仮説 } H_0 \quad : \quad \mu_B = 1180$$
$$\text{対立仮説 } H_1 \quad : \quad \mu_B < 1180$$

を $n = 100, \bar{x} = 1140$ に基づいて検定する.

【解法】　仮説 H_0 を認めた場合, 標本の結果が起こる確率がどのくらいかを計算して, あらかじめ与えた値より

$$\text{大} \quad \longrightarrow \quad \text{採択}$$
$$\text{小} \quad \longrightarrow \quad \text{棄却}$$

仮説を認めると電球 B の寿命 X の平均は 1180h, 標準偏差は 90h, この時, 中心極限定理を用いると $\overline{X}$ は平均 1180h, 標準偏差 $\dfrac{90}{\sqrt{100}} = 9h$ の正規分布で近似できる.　従って

$$\begin{aligned} P(\overline{X} \leqq 1140) &= P\left(\frac{\overline{X} - 1180}{9} \leqq \frac{1140 - 1180}{9}\right) \\ &= P\left(Z \leqq \frac{-40}{9}\right) \fallingdotseq P(Z \leqq -4.44) \\ &= P(Z \geqq 4.44) \quad \text{これは非常に小さい値} \end{aligned}$$

つまり仮説を仮定した場合, $\bar{x} \leq 1140$ になる事は常識ではありえない.

- もし標本平均 $\overline{X} = 1160$ であれば

$$\begin{aligned} P(\overline{x} \leqq 1160) &= P(Z \leqq \frac{-20}{9}) \\ &\fallingdotseq P(Z \leqq -2.22) &= 0.5\text{-}0.4868 \\ & &= 0.0132 \\ & &= 1.32\% \end{aligned}$$

これでもかなり怪しい.
(セールスマンが嘘をついているらしい)

- もし標本平均 $\overline{x} = 1170$ であれば

$$\begin{aligned} P(\overline{X} \leqq 1170) &= P(Z \leqq -1.11) \\ &\fallingdotseq 0.5\text{-}0.3665 \\ &= 0.1335 \\ &= 13.35\% \end{aligned}$$

このようなことはあり得る.

採択か棄却かを判定する基準 α を**有意水準**という.

α=0.05 の時

$\overline{x} = 1140$ の時 $\longrightarrow$ 棄却
$\overline{x} = 1160$ の時 $\longrightarrow$ $\alpha > 0.013$ より棄却
} A と B の寿命は異なると判定

$\overline{x} = 1170$ の時 $\longrightarrow \alpha < 0.1335$ より採択 A と B の寿命は異なると判断できない.

棄却：有意である　　採択：有意ではない

例のデータでは $\overline{x} = 1140$ であったので棄却. 従って B は A よりも劣っていると結論された.

α が与えられたとき,

$$\alpha = P(\overline{X} < x_\alpha | H_0 \text{ が真 })$$

となる値 x_α が定まる. $\overline{x} < x_\alpha$ を棄却域, x_α を棄却域の境界という.

上の例で $\alpha = 0.05$ の場合での $x_\alpha = x_{0.05}$ を計算してみる. 仮説が正しいとすると, 正規化（規準化）された値 $Z = -\dfrac{\overline{X} - 1180}{9}$ は（近似的に）標準正規分布に従うことがわかる. 正規分布表から $P(Z \geqq 1.645) = 0.05 = \alpha$ であるので（この場合 1.645 を正規化した棄却域の境界とよび $z_\alpha = z_{0.05}$ とかくことにする.）

$$\begin{aligned} x_{0.05} &= 1180 - 9 \times Z_0 = 1180 - 9 \times 1.645 \\ &= 1180 - 14.8 \\ &= 1165.2 \end{aligned}$$

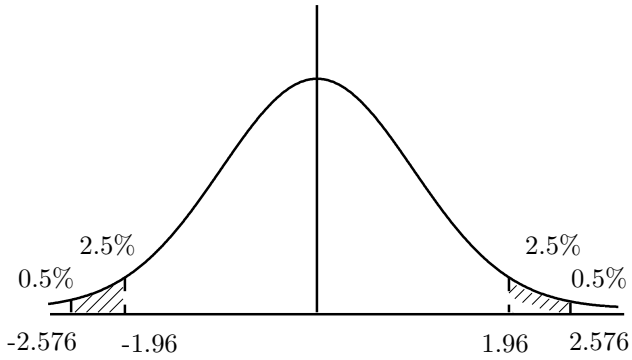
となる. 計算では, まず正規化（規準化）した値 $\overline{z}$ にたいして正規化した棄却域を用いると便利である.

$$|\overline{z}| > z_\alpha \Rightarrow \text{棄却}$$
$$|\overline{z}| \leqq z_\alpha \Rightarrow \text{採択}$$

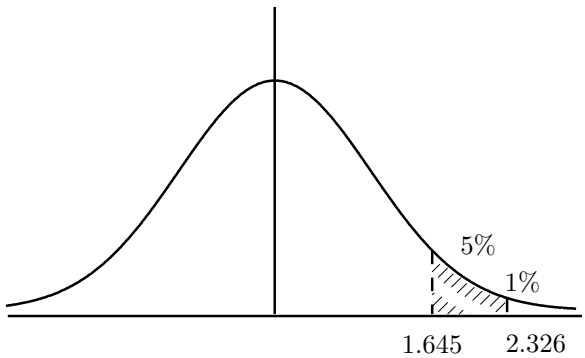
z_α は 片側検定の場合 $P(Z \geq z_\alpha) = \alpha$ となる正の値であり, 両側検定の場合 $P(|Z| \geq z_\alpha) = \alpha$ となる正の値である. $\alpha = 1\%, 5\%, 10\%$ の場合の z_α の値を表にしてみると, つぎのようになる.

α	z_α	
	両側検定	片側検定
0.01	2.576	2.326
0.05	1.96	1.645
0.10	1.645	1.286

上述の**例題 1**（片側検定） の場合では, $|\overline{z}| = |\frac{1140-1180}{9}| = 4.44$ であるので $\alpha = 1\%, 5\%, 10\%$ のいずれの場合も棄却であることが分かる.



両側検定



片側検定

例題 2 (両側検定) ある学校で過去数年間の新入生についての記録では適性検査の得点は平均 115, 標準偏差 20 であった. 今年の新入生に適性検査を行ったところ, 標本数 $n = 50$, $\overline{x} = 118$ を得た. 今年の新入生が今までの学生と比べて優れているか劣っているかを検定せよ. (ただし標準偏差は毎年同じであるとしてよい.)

解 標本値 $n = 50, \overline{X} = 118$ をもとに次の両側検定する ($\sigma = 20$ とする.)：

仮説 H_0 : $\mu = 115$(例年と同じ)
対立仮説 H_1 : $\mu \neq 115$(異なる)

H_0 の下で $\bar{X}$ は平均 $\mu_{\bar{X}} = 115$, 標準偏差 $\sigma_{\bar{X}} = \frac{20}{\sqrt{50}} = 2.83$ の正規分布で近似できる.

$\alpha = 0.05$ で両側検定すると, 正規化された棄却域は

$$|\overline{z}| > z_{0.05} = 1.960$$
$$|\overline{z}| = \left|\frac{118 - 115}{2.83}\right| = 1.06 < 1.960$$

従って H_0 は採択（有意ではない.）よって新入生は今までと異なると結論できない.

$\alpha = 0.1$ のときも $1.06 < z_{0.1} = 1.645$ より採択 (有意ではない).

仮説の採択について　有意性検定は, 仮説の下で我々が期待する結果が生じなかったことを根拠として, 仮説を棄却（つまり否定）することが主な内容である. これは論理学では **背理法**といわれているものである. あくまでも棄却されることが中心であって, 仮説が棄却されなかったからといって積極的に支持されたわけではない. 単に「結果が仮説と矛盾はしない」ことがいわれただけである. 仮説を採択したからといって仮説が真であることを積極的に「証明」したわけではない.

3.3.2　割合の検定

i 番目の試行を行ったとき

$$\begin{cases} \text{成功の場合} \longrightarrow & X_i = 1 \\ \text{失敗の場合} \longrightarrow & X_i = 0 \end{cases}$$

と X_i を定める.

n 回の試行のうち成功の回数 $\Sigma_n = \sum_{i=1}^{n} X_i$, 成功の確率 p、失敗の確率を $q = 1 - p$ とすると, Σ_n の分布は二項分布 $B_i(n, p)$ になる. 二項分布 $B_i(n, p)$ の平均は np, 標準偏差は $\sqrt{npq}$ であるから

$$\widehat{p} \equiv \quad \text{成功の割合} \equiv \frac{\Sigma_n}{n}$$

の平均は p、標準偏差は $\sqrt{\dfrac{pq}{n}}$ であることが分かる.

例題 3 飲酒運転率 (両側検定)　1974 年 7 月 11 日付の Los Angeles Times には, カリフォルニア大学医療心理学教授のアービン氏がオレンジ郡保健局と協力して, オレンジ郡の週末の運転者中の飲酒運転者の数を調べた調査報告が掲載されている. アメリカ全州の飲酒運転率は約 5% であるという. オレンジ郡で停車を命ぜられて検査を受けた 1000 人のドライバーのうちの 7% が飲酒運転であると判定された. 飲酒運転かどうかの判定は血液中のアルコールの量によってなされる. オレンジ郡のドライバーの真の飲酒運転率は全国のドライバーのそれと同じであるという仮説を, 同じでないという対立仮説に対して, $\alpha = 0.05$ で検定せよ.

解

カリフォルニア州の飲酒運転率 p とおく.

$$\left.\begin{array}{ll} \text{仮説 } H_0 & : p = 0.05 \\ \text{対立仮説 } H_1 & : p \neq 0.05 \end{array}\right\} \text{として仮説検定する.}$$

仮説 H_0 の下で n は十分大きいので
$\widehat{p}$ は平均 0.05, 標準偏差 $\sqrt{\dfrac{0.05 \times 0.95}{1000}} \fallingdotseq 0.0063$, である正規分布で近似される. $\alpha = 0.05$, 両側検定での棄却域は $|\overline{z}| > z_{0.05} = 1.960$ であるので

$$|\overline{z}| = \frac{0.07 - 0.05}{0.0063} = 2.9 > 1.960 \longrightarrow H_0 \text{は棄却}$$

よってカリフォルニア州の飲酒運転率は全州と異なると結論された.

3.3.3　2 つの平均値差の検定

2 つの母集団があり,

$$\begin{cases} \text{母集団}1: & \text{平均}\mu_1 \quad \text{標準偏差}\sigma_1 \\ \text{母集団}2: & \text{平均}\mu_2 \quad \text{標準偏差}\sigma_2 \end{cases}$$

であるとする. 母集団 i から大きさ n_i の標本をとり, 標本平均 $\overline{x}_i$, 標本標準偏差 s_i を得た. $(i = 1, 2.)$

- **近似定理**

  母集団分布が共に任意で母集団標準偏差 σ_1, σ_2 が未知のとき, 大きさ $n_1, n_2 (n_1, n_2 \geqq 50)$ の標本をとると, 2 つの平均の差 $\overline{X}_1 - \overline{X}_2$ の分布は

  $$\text{平均 } \mu_1 - \mu_2, \qquad \text{標準偏差 } \sqrt{s_1^2/n_1 + s_2^2/n_2}$$

  である正規分布で近似される.

例題 4 市役所の電球の購入

市役所は 2 つの銘柄の電球 1 , 2 のどちらかを購入する事にした. そこで電球☒, ☒の平均寿命を調べる.

$$\begin{array}{l} \text{電球 1 の } n_1 = 100 \text{ の標本から } \overline{x}_1 = 1160h, s_1 = 90h \\ \text{電球 2 の } n_2 = 100 \text{ の標本から } \overline{x}_2 = 1140h, s_2 = 80h \end{array}$$

を得た. 2 つの銘柄に差があるかどうか両側検定を行う.

電球☒, ☒の平均寿命をそれぞれ μ_1, μ_2 とおく. そして

$$\begin{cases} \text{仮説 } H_0 & \mu_1 = \mu_2 \quad (2 \text{ つの銘柄に品質の差は無い.}) \\ \text{対立仮説 } H_1 & \mu_1 \neq \mu_2 \quad (2 \text{ つの銘柄の品質に差がある.}) \end{cases}$$

とする.

H_0 の下で $\overline{X}_1 - \overline{X}_2$ の分布は平均 0, 標準偏差 $\sqrt{\dfrac{90^2}{100} + \dfrac{80^2}{100}} \fallingdotseq 12$ の正規分布で近似できる.

- $\alpha = 0.05$ で両側検定する.

  $$\begin{array}{c} \text{棄却域は } |\overline{z}| > z_{0.05} = 1.96 \\ \left| \dfrac{1160 - 1140}{12} \right| \fallingdotseq 1.67 < 1.96 \end{array}$$

  採択となる. 従って “2 つの銘柄に差があるとはいえない ” と結論される.

- $\alpha = 0.10$ で両側検定する.

  $$\begin{array}{c} \text{棄却域は } |\overline{z}| > z_{0.1} = 1.645 \\ \left| \dfrac{1160 - 1140}{12} \right| \fallingdotseq 1.67 > 1.645 \end{array}$$

  棄却となる. 従って “2 つの銘柄には差がある ” と結論される.

3.3.4 2 つの割合の差の検定

母集団 i では母集団の成功の割合が p_i である. $(i = 1, 2.)$ ここから n_i 個の標本をとる. この時標本からの成功の割合を $\widehat{p}_i$ と書く. つまり n_i 個の標本の中からの成功の個数（回数）を $\Sigma_{n_i}^{(i)}$ とおくと $\widehat{p}_i = \dfrac{\Sigma_{n_i}^{(i)}}{n_i}$ となる.

- **近似定理**

母集団☒と☒の成功の割合 p_1 と p_2 が共に未知のとき大きさ $n_1, n_2 (n_i \geqq 50, n_2 \geqq 50)$ の標本をとり, 標本からの成功の割合 $\widehat{p}_1, \widehat{p}_2$ を得た. 確率変数 $\widehat{p}_1 - \widehat{p}_2$ は

$$\text{平均：} p_1 - p_2, \qquad \text{標準偏差：} \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

の正規分布で近似できる. 特に仮説 $H_0 : p_1 = p_2 = p$ のもとでは

$$\widehat{p} = \frac{\Sigma_{n_1}^{(1)} + \Sigma_{n_2}^{(2)}}{n_1 + n_2}$$

を用いて

$$\text{平均：} 0 \qquad \text{標準偏差：} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n_1} + \frac{\widehat{p}(1-\widehat{p})}{n_2}} = \sqrt{\frac{(n_1+n_2)\widehat{p}(1-\widehat{p})}{n_1 n_2}}$$

とすることができる.

例題 5 ビタミン C の風邪に及ぼす効果 (片側検定)

ビタミン C の大量投与を受けた 407 人のうち, 期間中 105 人が風邪にかからず, 偽薬投与を受けた 411 人のうち 76 人が風邪にかからなかった. ビタミン C の風邪の予防効果を調べよ.

解

$$\begin{aligned} p_1 &: \text{ビタミン C 投与を受けた集団が風邪にかからない割合} \\ p_2 &: \text{偽薬投与を受けた集団が風邪にかからない割合} \end{aligned}$$

とおき,

$$\begin{cases} \text{仮説 } H_0 & : \ p_1 = p_2 (\text{ビタミン C は風邪に効果なし}) \\ \text{対立仮説 } H_1 & : \ p_1 > p_2 (\text{ビタミン C は風邪に効果あり}) \end{cases}$$

にたいして $\alpha = 0.05$ で片側検定する. 標本は

$$n_1 = 407,\ \Sigma_{n_1}^{(1)} = 105, \quad n_2 = 411,\ \Sigma_{n_2}^{(2)} = 76$$

であるので

$$\widehat{p}_1 = \frac{105}{407} \fallingdotseq 0.26, \quad \widehat{p}_2 = \frac{76}{411} \fallingdotseq 0.18, \quad \widehat{p} = \frac{\Sigma_{n_1}^{(1)} + \Sigma_{n_2}^{(2)}}{n_1 + n_2} = \frac{181}{818} \fallingdotseq 0.22$$

となる. 上の近似定理をもちいると $\widehat{p}_1 - \widehat{p}_2$ の分布は

$$\text{平均 } 0, \quad \text{標準偏差} \sqrt{\frac{0.22 \times 0.78}{407} + \frac{0.22 \times 0.78}{411}} = 0.029$$

の正規分布で近似される.

- $\alpha = 0.05$ で片側検定では棄却域は $|\overline{z}| > z_{0.05} = 1.645$ なので

$$\overline{z} = \left| \frac{0.26 - 0.18}{0.029} \right| \fallingdotseq 2.76 > 1.645$$

ゆえに H_0 は棄却. 従って ”ビタミン C は風邪の予防に効果がある ” と結論された.

3.3.5 検定の演習問題

(i) レンガの買い手は近頃のレンガの品質はどうも低下しているようだと思っていた. 過去の経験では, レンガの平均破壊強度は 400 ポンドで, 標準偏差は 20 ポンドであった. 100 個のレンガの標本をとってテストしたら, その平均は 390 ポンドであった. 平均品質は低下しているという対立仮説に対して, それは変わっていないという仮説を検定せよ.

(ii) 生物学者がある種の昆虫の 50% が殺せるように殺虫液を調合した. この液をこの種の昆虫 200 匹に吹きかけたところ, 120 匹が死んだ. 液の調合はうまくいったと結論してよいか.

(iii) 50 人ずつの小学校の生徒 2 組に 2 通りの違った方法で読み方を教えた. 授業後読み方の試験を行い, 次の結果を得た. $\overline{x}_1 = 73.4, \overline{x}_2 = 70.2, s_1 = 9, s_2 = 10$. この結果から仮説: $\mu_1 = \mu_2$ を検定せよ.

(iv) ある都市におけるテレビの視聴率調査で, ある番組を男子は 200 人中 56 人が好まなかったのに対し, 女子は 300 人中 75 人がそれを好まなかった. 両者の嗜好に実際上差があるといえるだろうか.

(v) x は未知の平均 μ, 標準偏差 $\sigma = 6$ の正規分布に従うとする. 大きさ 16 の標本をとり, $\overline{x} = 33$ を得たとき,

(a) 仮説 $H_0 : \mu = 30$ を, 対立仮説 $H_1 : \mu > 30$ に対して検定せよ.

(b) しばらくしてから, 大きさ 32 の第 2 の標本をとり, $\overline{x} = 29$ を得た. 母集団平均 μ がこの間に変化したとみなせるか. 平均は変化していないという仮説を, 平均は小さくなったという対立仮説に対して検定せよ.

(vi) ある市で 200 人の自動車所有者の標本を調べたら, そのうち 48 人が期限切れの運転免許証を持っていた. 前年度の期限切れ運転免許証所持率は 0.30 であった. これらのデータを用いて、次の問題を解け.

(a) 母集団割合は $p = 0.30$ であるとの仮説を検定せよ.

(b) 母集団割合は $p = 0.30$ であるとして, 大きさ 50 の標本を毎日とる場合の $\widehat{p}$ に対する管理図の管理限界を求めよ.

(vii) 1974 年 4 月 27 日付の Newsweek に報道された実験は, マリファナが性活動に及ぼす影響を調べたもので, セントルイスの生殖研究協会によって行われた. 実験には 20 人の健康な若者が選ばれ, 1 週間に少なくとも 4 日, 最低 6 週間にわたりマリファナを吸わせたが, この期間中他の薬は一切使用させなかった. 対照群として, マリファナを吸わない 20 人の若者が比較のために選ばれた. 使用した性活動の尺度は, 血液中の男性ホルモン “テストステロン” の量であった. 添字の 1 と 2 はそれぞれマリファナ群と非マリファナ群に対応するとし, 実験の結果, 次のデータを得たとする.
$\overline{x}_1 = 416, \overline{x}_2 = 742, s_1 = 152, s_2 = 130.$

(a) 仮説 $H_0 : \mu_1 = \mu_2$ を検定せよ.

(b) 同じ仮説を小標本法によって検定せよ.

(viii) 1973 年 10 月 22 日付の Time の報道によると, 酒を飲めば顔が赤くなるという東洋人の主張が真実か否かを調べる研究がノースカロライナ州のチャペルヒルで行われた. この研究では適度な飲酒習慣をもつ東洋系アメリカ人 24 人と同じ程度の飲酒習慣をもつヨーロッパ系アメリカ人 24 人とを比較した. 結果は次のようであった.

東洋系アメリカ人は 24 人中 17 人が赤くなった.

一方, ヨーロッパ系アメリカ人は 24 人中 3 人が赤くなった.

仮説 $H_0 : p_1 = p_2$ を検定せよ. ここで p_1, p_2 は両グループにおける割合を表す.

(ix) 以下の数値は $\mu = 3.0, \sigma = 0.5$ の正規母集団からの標本抽出によって得られたものである. おのおの 25 個の標本が 2 人の実験者 A, B によってとられた. これらのデータを用いて, 次の問題を解け.

(a) 全データを 1 組のデータとみなして, 仮説 $H_0 : \mu = 3.0$ を対立仮説 $H_1 : \mu \neq 3.0$ に対して検定せよ.

(b) これら 50 個の観測値から管理図を作れ.

(c) $\sigma_A = \sigma_B = \sigma = 0.5$ を用いて, $\mu_A = \mu_B$ なる仮説を検定せよ.

(d) 標本偏差が既知ではないとして (c) を解け.

(e) A のデータと t 分布を用いて, $H_0 : \mu_A = 3.3$ を $H_1 : \mu < 3.3$ に対して検定せよ.

(f) t 分布を用いて, 仮説 $H_0 : \mu_A = \mu_B$ を検定し, このときの t の値を (d) の z の値と比べよ.

A	3.5	3.4	2.8	2.6	2.3	2.4	2.9	2.9	3.1	3.3	3.9	3.0	2.9
B	2.9	3.6	3.1	2.5	2.6	3.6	2.9	2.1	2.6	2.5	2.3	2.6	3.3

3.0	2.5	2.7	3.9	2.6	2.8	3.2	3.6	2.7	3.0	3.8	3.1
3.9	2.7	2.4	2.8	3.0	3.3	3.7	3.3	3.1	2.7	2.4	3.3

参考文献

[1] 初等統計学：ポール G. ホーエル 著, 浅井 晃, 村上 正康 訳, 培風館.

[2] 統計解析入門：篠崎信雄 著, サイエンス社.

[3] やさしい統計学入門: 田栗正章, 藤越康祝, 柳井晴夫, C.R. ラオ 著, 講談社.

[4] やさしい統計: 吉原健一, 金川秀也 著, 培風館.

[5] 統計学入門：基礎統計学☒, 東京大学出版会.