統計学入門

東京国際大学教授 張本 浩

目　次

第1章　平均値と分散

1.1 平均値(mean)

表 1.1：野球選手 16 人の身長

番号	身長(x_i)	番号	身長(x_i)
1	179	9	183
2	181	10	174
3	176	11	180
4	165	12	182
5	173	13	179
6	173	14	171
7	178	15	182
8	178	16	179

平均身長：16 人の身長をすべて足して 16 で割った数値

(179+181+176+・・・+182+179)÷16=177.06

＊ 観察データ：$x_1, x_2, \cdots, x_n$ あるいは $x_i (i = 1, 2, \cdots, n)$

＊ 平均値：一群のデータをただ 1 つの数値で表わす指標(=代表値)

[諸平均値]

・算術平均(arithmetic mean)

それぞれのデータに同じ重みを与えて求めた平均

・加重平均(weighted mean)

それぞれのデータに異なる重みを与えて求めた平均

・幾何平均(geometric mean)

比率データに適用される平均

・調和平均(harmonic mean)

平均速度を求める場合に用いられる平均

・RMS(Root Mean Square)：データの正負に意味のない場合に用いられる平均

$$\mathrm{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

(1)算術平均

$$\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1.1)$$

[算術平均の３つの性質]

・「算術平均はデータの１次変換を保持する」

１次変換：元のデータ(x_i)と任意の定数$(a\ \&\ b)$を用いて別のデータ(y_i)を作ること

$$y_i = ax_i + b \qquad (1.2)$$

**１次変換前後のデータの関係

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) = \frac{a}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} b$$

$$= a\bar{x} + \left(\frac{1}{n}\right)nb = a\bar{x} + b \qquad (1.3)$$

(1.3)式から

$$\bar{x} = \frac{\bar{y} - b}{a} \qquad (1.4)$$

[例題１]１次変換と平均値の計算

電球の寿命(x_i)	1812	1792	1795	1804	1810
１次変換値(y_i)					

・「偏差(データと算術平均との差)の和はゼロである。」

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0 \qquad (1.5)$$

・「算術平均は偏差の平方和を最小にする値である。」

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \le \sum_{i=1}^{n}(x_i - a)^2 \qquad (1.6)$$

証明：$S(a)=\sum_{i=1}^{n}(x_i-a)^2$とし，$S(a)$を最小にする$a$の値を求めるために

$$\frac{dS(a)}{da}=-2\sum(x_i-a)=0 \ \rightarrow \ a=\bar{x} \qquad (1.7)$$

(2)加重平均(weighted mean)

・各データ(x_i)の重み(w_i)がそれぞれ異なる場合に用いられる平均

$$\bar{x}_w=\frac{\sum_i W_i x_i}{\sum_i w_i} \qquad (1.8)$$

［例題 2］年度別売上高の算術平均値と加重平均値

年　　度	2016	2017	2018	2019	2020
売上高(x_i)	14	18	20	25	28
重　み(w_i)	1.0	1.1	1.2	1.5	2.0

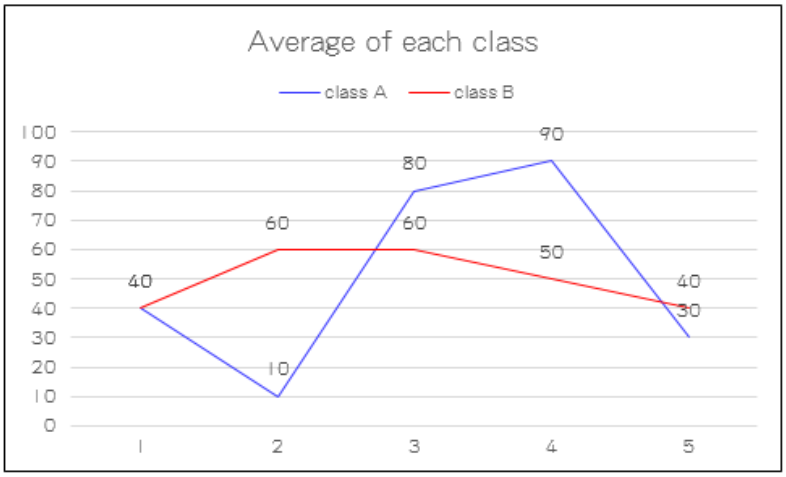
(3)中央値(median)

データを大きさの順に並べたときちょうど中央に位置するデータの値を中央値という。中央値はデータの数(n)が奇数のときは$(n+1)/2$番目の値であり，nが偶数のときは$n/2$番目と$\left(\frac{n}{2}\right)+1$番目のデータの中間値になるが，便宜的にその２つの値の算術平均とするのがふつうである。

1.2 分散度の指標

[例題 3] A クラス(5 人)の試験成績：40, 10, 80, 90, 30 → $\bar{x}_A = 50$

B クラス(5 人)の試験成績：40, 60, 60, 50, 40 → $\bar{x}_B = 50$



“A クラスと B クラスのデータの散らばりの度合い(=分散度)には明らかな差が存在する。”

(1)範囲と四分位偏差

・範囲(range)：$R = x_{max} - x_{min}$ (1.9)

・四分位偏差(Quartile Deviation)：$QD = \frac{Q_3 - Q_1}{2}$ (1.10)

・範囲と四分位偏差はデータの一部を用いてデータの分散度を示そうとする指標であるので，あまり実用的ではない。

[例題 4] 次のデータの範囲と四分位偏差を求めよ。

No.	1	2	3	4	5	6	7	8	9	10	11	12
data	30	49	84	29	47	56	81	12	90	59	66	3

(2)平均絶対偏差(Mean Absolute Deviation)：偏差の絶対値の算術平均

$$MAD = \frac{1}{n}\sum_{i=1}^{n} | x_i - \bar{x} | \quad (1.11)$$

[例題 5] 例題 3 の A クラスと B クラスの平均絶対偏差を求めよ。

(3)標準偏差(standard deviation)

データの分散度を測るもっとも一般的な指標であり，次のように定義される。

「 **偏差(データと平均値との差)の２乗値を算術平均し，**
それを元のデータと同次元に戻すために
平方に開いたもの 」

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (1.12)$$

標準偏差は常に正の値であり，標準偏差が大きいほど分散度も大きい。

標準偏差の２乗値を分散(variance)という。

・分散の簡便計算式

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n}x_i + \frac{1}{n}n\bar{x}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\bar{x}^2 + \bar{x}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2 \qquad (1.13)$$

［例題 6］ 例題３の Ａクラスと Ｂクラスの標準偏差を求めよ。

1.3 標準偏差の性質

(1)データの一次変換ど標準偏差

元のデータ(x_i)を 1 次変換して得られる別のデータ(y_i)の分散は次式により求められる。

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}[(ax_i + b) - (a\bar{x} + b)]^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}a^2(x_i - \bar{x})^2 = a^2 s_x^2 \qquad (1.14)$$

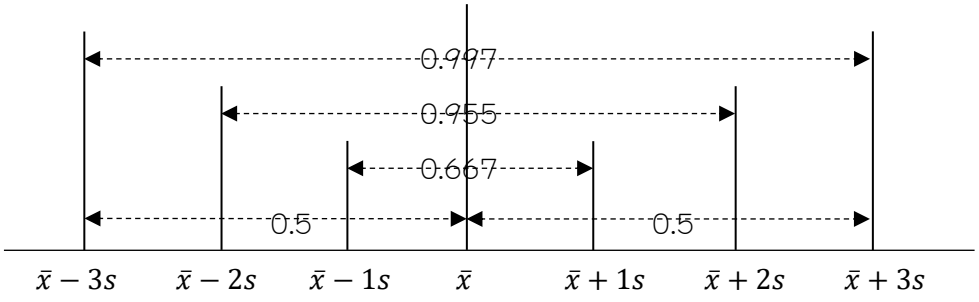
(1.14)式から　$s_y = |a|\, s_x$　　(1.15)

(1.15)式から　$s_x = \frac{s_y}{|a|}$　　(1.16)

[例題 7] 1 次変換を利用してデータ(x_i)の平均値と標準偏差を求めよ。

データ (x_i)	280	330	310	290	320
1 次変換値 (y_i)					

(2)平均・標準偏差どデータの割合との対応

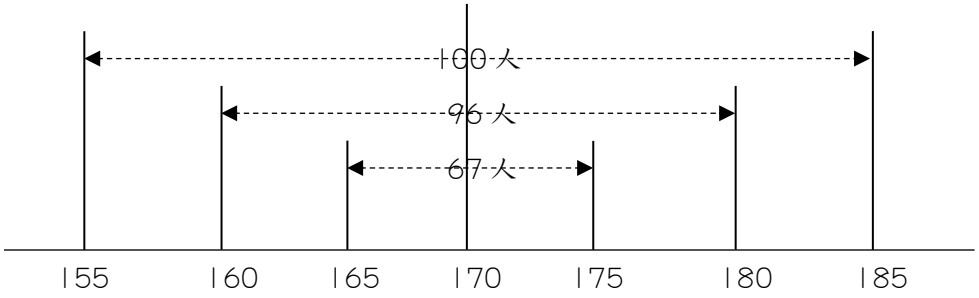


$\Pr(\bar{x}-1s < x < \bar{x}+1s) = 0.667$

$\Pr(\bar{x}-2s < x < \bar{x}+2s) = 0.955$ & $\Pr(x > \bar{x}) = \Pr(x < \bar{x}) = 0.5$

$\Pr(\bar{x}-3s < x < \bar{x}+3s) = 0.997$

例えば，東京都内の高校 3 年生 100 人の身長を測り，その平均値と標準偏差を求めて見たところ，平均値が 170cm，標準偏差が 5cm であった場合，データの範囲と人数は次のように表わせる。



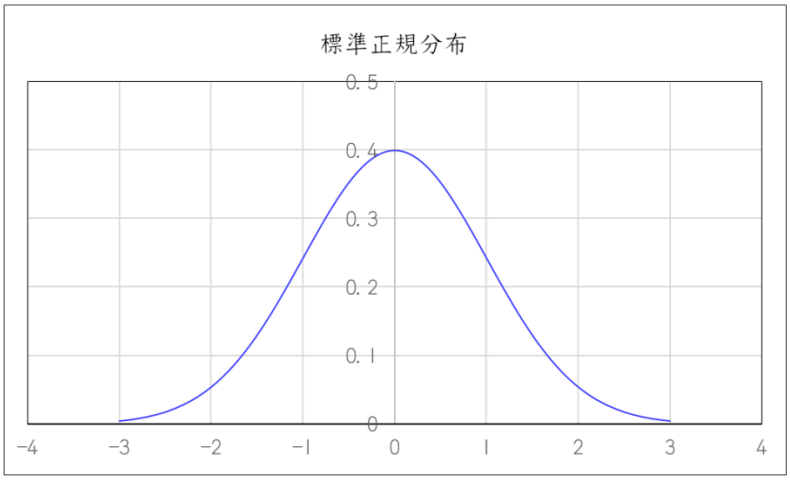
1.4 標準化変量と偏差値

(1)標準化変量

平均と標準偏差の異なるデータ群を直接比較できるように調整した数値

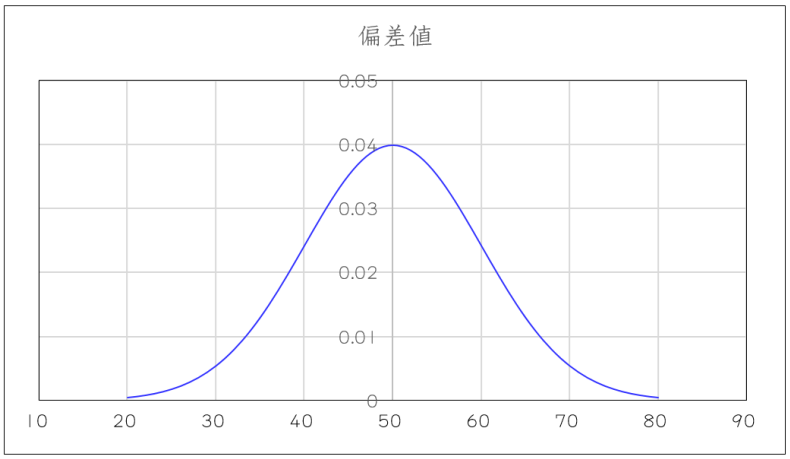
元のデータ(x_i) ⟶(標準化)⟶ 標準化変量(z_i)

$(\bar{x}, s_x)$ $\boxed{z = \frac{x_i - \bar{x}}{s}}$ (1.17) $(\bar{z} = 0, s_z = 1)$



(2)偏差値(standard score)：平均 50，標準偏差 10 となるように調整した値

$$ss_i = 50 + 10 \times \left(\frac{x_i - \bar{x}}{s}\right) \quad (1.18)$$



例えば，5 個のデータ(85, 46, 62, 28, 79)の平均値と標準偏差および 1 番目のデータ(85)の偏差値は次のように求められる

$$\bar{x} = \frac{85+46+62+28+79}{5} = 60$$

$$s = \sqrt{\frac{85^2+46^2+62^2+28^2+79^2}{5} - 60^2} = 21.02$$

$$ss_1 = 50 + 10 \times \frac{85-60}{21.02} = 61.89$$

[練習問題 1]

ある学生が英語の試験を 3 回受けて 90 点, 80 点, 70 点の成績を収めた。担当教師は，最終の成績を決めるにあたって，2 回目の試験は 1 回目の試験の 3 倍の重要さをもち，3 回目の試験は 2 回目の試験の 2 倍の重要さをもつと考えている。この学生の英語の平均点はいくらになるか。

[練習問題 2]

数学の試験で，630 人の学生の平均点は 69.6 点，標準偏差は 8.2 点であった。86 点以上の点数の学生はおよそ何人いるか。

[練習問題 3]

ある英語と数学の試験で，A 君はそれぞれ 72 点, 57 点という得点であった。クラスでの英語の平均点は 70 点，標準偏差は 5 点，数学の平均点は 50 点，標準偏差は 8 点であったとすれば，A 君にとって，相対的に(クラスの中の順位では)どちらの科目の方がよくできたといえるか。

第２章 度数分布

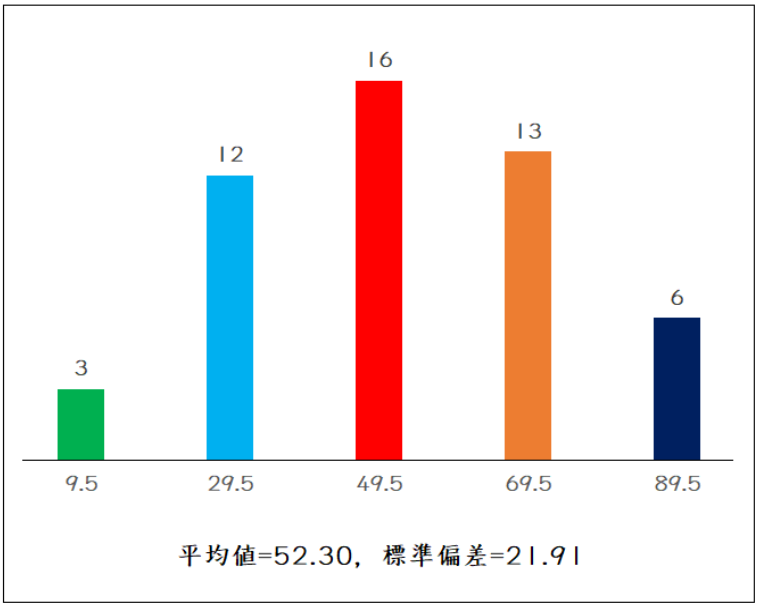
2.1 度数分布(frequency distribution)とは

データを大きさによっていくつかの級(class)に分けたとき，各級に属するデータの数(度数：frequency)を表や図に示したもの

33, 54, 67, 84, 25, 76, 93, 45, 47, 75, 39, 44, 57, 62, 73, 55, 27, 68, 36, 83, 58, 95, 35, 77, 52, 29, 15, 87, 42, 58, 34, 72, 25, 61, 50, 71, 13, 47, 56, 21, 64, 26, 81, 48, 43, 23, 65, 41, 6, 78

＜度数分布表＞

級	0－19	20－39	40－59	60－79	80－99	計
中央値	9.5	29.5	49.5	69.5	89.5	－
度　数	3	12	16	13	6	50



＊1 群のデータの分布状況(全体的特徴)を把握するために作成するのが度数分布

＊ データをいくつかの級(class)に分け，各級に入るデータの数(度数)を示したもの

＊ 度数分布≡データの要約→全体的情報の獲得 vs.詳細情報の流失

＊ 情報の流失度は級の数に反比例する(級の数が少なくなるほど, 情報流失度が高くなる)。

＊ 級の数(間隔)の決定要因：データの大きさ，分析者の求める情報の質，等々。

2.2 度数分布のグラフ

・ヒストグラム(histogram)：度数の大きさを柱の高さで表したグラフ

・相対度数：(各級の度数÷全体度数)×100

・累積度数：各級の度数を累積して得られる度数

・中央値：(下限値＋上限値)÷2

[例題 1] テスト点数の分布

級(下限値)	級(上限値)	度数(人数)	相対度数(%)	累積度数(人数)	中央値(x_k)
00	09	1			
10	19	0			
20	29	4			
30	39	6			
40	49	8			
50	59	8			
60	69	10			
70	79	7			
80	89	4			
90	99	2			
計		50		＊	＊

2.3 度数分布の作り方

(1) 級の数を適切に決める。

(2) 級の間隔は均一にする。

(3) 級の限界を明確にする。

(4) 度数の集中点がある場合、その点が級の中央にくるようにする。

＊スタージスの公式：級の間隔(c)を決めるための式

c：級の間隔，R：範囲，m：級の数，n：データの数

$$m = 1 + \frac{\log_{10} n}{\log_{10} 2} = 1 + \frac{\log_{10} n}{0.301} = 1 + 3.32\, log_{10} n \qquad (2.1)$$

$$\therefore \mathrm{c} = \frac{R}{m} = \frac{x_{max} - x_{min}}{1 + 3.32 \log_{10} n} \qquad (2.2)$$

[例題 2] 度数分布表の作成

20 問の試験で50 人の受験者が正解した問題数				
13	9	5	11	14
6	5	8	11	13
10	16	15	3	19
18	9	9	5	12
13	12	15	9	18
12	16	7	12	13
11	18	15	9	21
9	11	6	12	12
10	16	2	14	10
17	8	15	11	12

2.4 度数分布からの平均値・分散の計算

各級の中央値(x_k)，各級の度数(f_k)，級の数(m)，度数の合計値(n)

(1)平均値

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \cdots + f_mx_m}{f_1 + f_2 + \cdots + f_m} = \frac{1}{n}\sum_{k=1}^{m} f_kx_k \qquad (2.3)$$

(2)分散

$$s^2 = \frac{1}{n}\sum_{k=1}^{m} f_k(x_k - \bar{x})^2 \qquad (2.4)$$

$$s^2 = \frac{1}{n}\sum_{k=1}^{m} f_kx_k^2 - \bar{x}^2 \qquad (2.5)$$ ←分散の簡便計算式

[例題 3] 度数分布の平均値と分散

中央値(x)	10	30	50	70	90	計
度　数(f)	17	45	53	27	8	150

[例題 4] 度数分布表の作成・度数分布の平均値と分散

パートタイマ 30 人の月間労働時間					
72	71	92	64	87	77
85	91	72	70	67	78
74	83	78	77	80	61
78	88	81	77	75	82
68	69	82	82	69	70

[練習問題 1]

2 個のサイコロを同時に投げるとき，6 の目が何個出るかを考えると，0 か 1 か 2 のいずれかである。2 個のサイコロを同時に投げる実験を 100 回行ない，6 の目が出た個数が 0，1，2 であったのがそれぞれ何回であったかを記録し，そのヒストグラムを描け。

[練習問題 2]

下の表はある 1 年間の 80 種類の月刊誌の発行部数の分布である。この分布のヒストグラムと折れ線グラフを描け。

発行部数(単位:千部)	月刊誌数	発行部数(単位:千部)	月刊誌数
100～199	1	600～649	5
200～299	1	650～699	7
300～399	3	700～749	30
400～400	4	750～799	18
500～599	3	800～849	8

[練習問題 3]

下記データの度数分布表を作成し，その度数分布の平均値と分散を求めよ。

求人倍率				
13.2	12.9	8.9	22.7	16.5
13.2	14.0	19.4	14.7	23.3
17.8	18.0	15.6	24.5	11.0
14.1	16.4	10.5	9.2	17.8
19.7	20.5	11.8	13.6	24.7
15.5	14.7	15.7	17.9	9.4
14.9	12.1	12.5	10.1	14.3
10.2	12.8	8.8	11.3	12.4
9.2	15.3	13.1	19.5	8.9
12.5	7.2	10.1	16.8	9.8

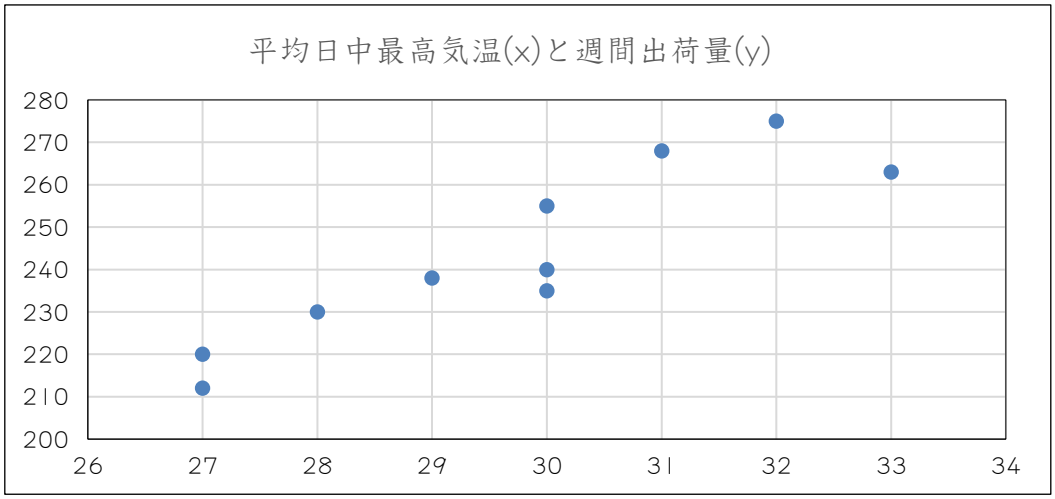
3.1 回帰関係の意味

表 3.1 は，ある年の夏の 10 週間について，ある清涼飲料会社の営業所の週間出荷量(y_i)とその週の平均日中最高気温(x_i)とを調べてものである。

これを見ると，平均日中最高気温が高い週には出荷量も概して多く，気温が低い週には出荷量が概して少ないことがわかる。

表 3.1 清涼飲料出荷量と気温

平均日中最高気温(x_i)	29	27	30	31	32	33	30	30	28	27
週間出荷量(y_i)	238	220	255	268	275	263	240	235	230	212



以上のように，回帰関係は変数 x が変数 y の平均値を決めるという関係であり，変数間の相関関係を記述・分析する方法が回帰分析(regression)である。回帰分析では，変数 x は回帰変数(独立変数=原因変数)，変数 y は被回帰変数(従属変数=結果変数)と呼ばれる。

・回帰分析の変数

変数 x：回帰変数(regressor) or 独立変数→原因の変数

変数 y：被回帰変数(regressand) or 従属変数→結果の変数

・回帰分析の種類

(1)独立変数の数による分類

単回帰(single regression)：独立変数が 1 つだけ場合の回帰

重回帰(multiple regression)：独立変数が 2 つ以上の場合の回帰

(2)独立変数の次数による分類

線形回帰(linear regression)：独立変数の次数が 1 の場合の回帰

非線形回帰(non-linear regression)：独立変数の次数が 2 以上の場合の回帰

・線形単回帰：2 変数間の相関関係を線形式(1 次式)で表わす。

$$\hat{y}_i = a + bx_i \quad (3.1) \rightarrow \text{線形単回帰式}$$

$\hat{y}$：従属変数の予測値，　a, b：推定係数

3.2 最小 2 乗法(Least Square Method)

・推定係数の推定(回帰式の決定)のための方法

> 従属変数の実測値と予測値との差(残差)の 2 乗の総計(残差平方和)が最小となるように推定係数$(a \,\&\, b)$を決める。

・残差平方和：$S(a, b)$

$$S(a, b) = \sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n}\{(y_i - (a + bx_i)\}^2 \quad (3.2)$$

$S(a, b)$がもっとも小さくなるように a と b の値を決めるために，(3.2)を a と b のそれぞれについて偏微分しそれらをゼロとおく。

$$\left.\begin{aligned} \frac{\partial S(a, b)}{\partial a} &= -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \\ \frac{\partial S(a, b)}{\partial b} &= -2\sum_{i=1}^{n}x_i(y_i - a - bx_i) = 0 \end{aligned}\right\} \quad (3.3)$$

(3.3)を整理すると，いわゆる正規方程式というものが得られる。

$$\left.\begin{aligned} na + b\sum x = \sum y \\ a\sum x + b\sum x^2 = \sum xy \end{aligned}\right\} \quad (3.4)$$

(3.4)の２式を連立して解けば，推定係数が求められる次式が得られる。

$$a = \frac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - (\sum x)^2} \quad , \quad b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \qquad (3.5)$$

[例題１]ある中古車の使用年数(x年)と販売価格(y万円)から線形回帰式($\hat{y}_i = a + bx_i$)を求めよ。

中古車の使用年数(x_i)	1	4	10	2	5	6	8	1
販売価格(y_i)	64	35	11	47	27	36	30	59

3.3 決定係数

線形回帰式の説明力を示す指標であり，次の２つの分散に基づいて求められる。

① 実測値の分散：$s_y{}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ (3.6)

② 回帰線のまわりにおける実測値の分散：$s_{yx}{}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2$ (3.7)

$$d = 1 - \frac{s_{yx}{}^2}{s_y{}^2} \qquad (3.8)$$

・完全相関($y = \hat{y}$)の場合，$s_{yx}{}^2 = 0 \rightarrow \therefore d = 1$

・実測値と予測値がかけ離れるほどs_{yx}の値は大きくなるので，d は小さくなる。

・不完全相関の場合，$s_{yx} = s_y \rightarrow \therefore d = 0$

・実測値が回帰線の近くに分布すればするほど，d は大きくなる。

[例題 2]下の表は，12 人の女性の年齢と血圧を示したものである。

年齢(x_i)	56	42	72	36	63	47	55	49	38	42	68	60
血圧(y_i)	147	125	160	118	149	128	150	145	115	140	152	155

年齢(x_i)に対する血圧(y_i)の線形回帰式と決定係数を求めよ。

[練習問題 1]

次に示すのは，10 の小さな宝石店について広告費支出(総経費に対する百分比)と純利益(売上高に対する百分比)を調べてものである。

広告費(x_i)	1.2	0.7	1.5	1.8	0.5	3.4	1	3	2.8	2.5
利益(y_i)	2.7	2.4	2.7	3.3	1.1	5.8	2.2	4.2	4.4	3.8

(a) このデータをプロットせよ。

(b) 広告費(x_i)に対する利益(y_i)の線形回帰式を求め，(a)で作られたグラフ上にその直線を描け。

(c) 広告費用が$x = 2.0\%$であるような店の純利益を推定せよ。

[練習問題 2]

ある講習会でこの講習を受けた 10 人の学生たちが全てのコースを終えるのに要した時間数と，最後に行ったテストの結果(点数)である。

要した時間(x_i)	30	25	50	38	20	70	35	24	60	45
テストの点数(y_i)	80	80	45	70	95	20	50	90	25	50

(a) このデータをプロットせよ。

(b) これらのデータに回帰直線をあてはめ，(a)で得られたクラフ上にその直線をプロットせよ。

(c) 決定係数を求めよ。

第４章 確率

4.1 順列と組合せ

(1)順列(permutation)：データの並べ方

いま３文字(a, b, c)を１列に並べると，その並べ方は次のように決められる。

３文字(a, b, c)の順列：(a, b, c) (a, c, b) (b, a, c) (b, c, a) (c, a, b) (c, b, a)

すなわち，異なる３文字の並べ方の数は $3 \times 2 \times 1 = 6$ 通りである。

一般に n 個の異なるものを並べる並べ方は

$$n(n-1)(n-2)\cdots 3 \times 2 \times 1 = n! \qquad (4.1)$$

通りある。 n! は nの階乗(factorial)という。

$$0! = 1, \quad 1! = 1, \quad 2! = 2, \quad 3! = 6, \quad 4! = 24, \cdots\cdots$$

である。

次に，今度は４文字(a, b, c, d)があるとして，これらの中から２つの文字を選んで並べる場合の並べ方は，次のようになる。

(a, b) (a, c) (a, d) (b, a) (b, c) (b, d) (c, a) (c, b) (c, d) (d, a) (d, b) (d, c)

一般に n 個の異なるものの中から $r(r \leq n)$個を選んで並べる並べ方は

$$n(n-1)(n-2)\cdots(n-r+1) \qquad (4.2)$$

通りある。なおこれは n!/(n − r)! に等しい。これを ${}_nP_r$ と書き，n 個の異なるものの中から r個をとる順列という。順列の数は次式により求められる。

$$\begin{aligned} {}_nP_r &= n(n-1)(n-2)\cdots(n-r+1) \\ &= \frac{n(n-1)(n-2)\cdots(n-r+1)(n-r)(n-r-1)\cdots 3 \times 2 \times 1}{(n-r)(n-r-1)\cdots 3 \times 2 \times 1} \\ &= \frac{n!}{(n-r)!} \qquad (4.3) \end{aligned}$$

[例題 1] P, Q, R, S, T の５人を１列に並べる。

(1) P を先頭に並べるとすると、その並べ方は何通りか。

(2) Q と R の２人を続けて並べるとすると、その並べ方は何通りか。

最後に，文字列のなかに同じ文字が混ざっている場合の順列の数は，例えば，TOYOTA の６文字の順列の数を求める場合，すべての文字数 $(n) = 6$ ，T が２つ$(n_1 = 2)$，O が２つ$(n_2 = 2)$であるので，次のように順列の数を求めることができる。

$$\frac{n!}{n_1! \times n_2!} = \frac{6!}{2! \times 2!} = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{(2 \times 1) \times (2 \times 1)} = 180$$

[例題 2] YOKOHAMA の８文字の順列の数を求めよ。

(2)組合せ(combination)：データの選び方

次に上の問題で，選び出された２文字の並べ方は問わず，どの２文字の組が選び出されたかだけを問題にする。すなわち，４文字(a, b, c, d)のなかから２文字を選び出す方法は何通りあるかということである。今度は文字を並べる順序は問わないので，上に列挙した12通りのうち，abとba, acとca,…などの各２通りはそれぞれ同じ２文字の組であるから区別されない。そこで答えは $\frac{12}{2} = 6$ 通りである。

$$(a, b)\ (a, c)\ (a, d)\ (b, c)\ (b, d)\ (c, d)$$

一般に n 個の異なるものの中から r個を選ぶ選び方は ${}_nP_r / r!$ である。このように n 個の異なるものの中から r個を選ぶ選び方をn 個の異なるものの中から r個をとる組合せといい，${}_nC_r$ という記号で表す。

$$ {}_nC_r = \frac{{}_nP_r}{r!} = \frac{n(n-1)(n-2)\cdots(n-r+1)}{r!} = \frac{n!}{r!\,(n-r)!} \qquad (4.4)$$

$$ {}_{10}C_3 = \frac{10!}{3!\,7!} = \frac{10 \times 9 \times 8}{3 \times 2 \times 1} = 120 $$

$$ {}_{10}C_7 = \frac{10!}{7!\,3!} = \frac{10 \times 9 \times 8}{3 \times 2 \times 1} = 120 $$

$$\therefore\ {}_nC_r = {}_nC_{n-r}$$

[例題 3] 女性４人，男性６人の中から代表を３人選びたい。女性が２人以上含まれる選び方は何通りあるか。

4.2 確率(probability)

・ある事象のおこる確からしさの度合い

・確率に対するアプローチ

①先験的確率(a priori probability)

正しいサイコロ投げの場合，各目の出る確率は予め決まっている。

∴ 事象 A の発生確率：$Pr(A) = \frac{n_A}{n}$

②経験的確率(empirical probability)

歪んだサイコロ投げの場合，ある目の出る確率は数多く観察しなければ

決められない。∴ 事象 A の発生確率：$Pr(A) = \lim_{n\to\infty}\left(\frac{n_A}{n}\right)$

③主観的確率(subjective probability)

1 回限りの賭けや大学受験などのように反復観察の不可能な場合の確率は意思決定者の主観(信念，確信，等々)によって決められるしかない。

④公理的確率(mathematical probability)

数学的概念としての確率は以下の 3 つの公理を満たさなければならない。

＜公理 1＞ $0 \le Pr(A) \le 1$

＜公理 2＞ $Pr(A) + Pr(B) + \cdots + Pr(Z) = 1.0$

＜公理 3＞ $A, B, \cdots Z$ が互いに排反事象の場合，

$$Pr(A \cup B \cup \cdots \cup Z) = Pr(A) + Pr(B) + \cdots + Pr(Z)$$

4.3 標本空間(sample space)

・実験(観察)の結果(=標本点)の集合

例えば，2 枚の硬貨なげ実験の場合，硬貨の「表=Head，裏=Tail」とすると，

標本空間：S={(Head & Head), (Head & Tail), (Tail & Head), (Tail &Tail)}

・硬貨 1 枚とサイコロ 1 個を投げる実験の標本空間

S={(Tail & 1), (Tail & 2), (Tail & 3), (Tail & 4), (Tail & 5), (Tail & 6),
(Head & 1), (Head & 2), (Head & 3), (Head & 4), (Head & 5), (Head & 6)}

4.4 標本点(sample point)と確率

・標本空間の各標本点にはそれぞれ確率が付与される。

(例 1)2 枚の硬貨なげ実験の場合：Pr(HH)=Pr(HT)=Pr(TH)=Pr(TT)=1/4

(例 2)硬貨 1 枚とサイコロ 1 個を投げる実験の場合：Pr(T1)=…=Pr(H6)=1/12

(例 3)サイコロを 2 回投げる場合の標本空間

＼	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

① 出た目の和が 10 以下である確率=33/36

② どちらか一方が 3 以下の目である確率=18/36

③ 2 つのサイコロの目が同じでない確率=30/36

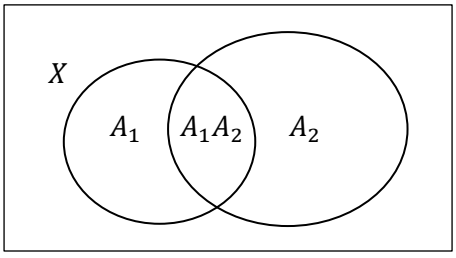
・余事象(complementary event)：

事象 A の余事象を$\overline{\mathrm{A}}$とすると，$Pr(\bar{A}) = 1 - Pr(A)$

・標本空間(S)の確率：$Pr(S) = Pr(A) + Pr(\bar{A})$

4.5 加法定理

・2 つの事象$(A_1 \& A_2)$がある場合



X：全体集合(標本空間)

$A_1\ or\ A_2$： $A_1 + A_2$ → 和事象

$A_1 \& A_2$： A_1A_2 → 積事象

加法定理：$Pr(A_1 + A_2) = Pr(A_1) + Pr(A_2) - Pr(A_1A_2)$ (4.5)

[例題 4] 1 から 10 までの数字の書かれた 10 枚のカードから 1 枚を抜くとき,

事象A_1：2 で割り切れる数，事象A_2：3 で割り切れる数，とすると,

和事象$(A_1 + A_2)$：2 あるいは 3 で割り切れる数

積事象(A_1A_2)：2 でも 3 でも割り切れる数

となる。各事象の確率を直接求めた後，(4.5)式により和事象の確率を求めることによって(4.5)式が正しいことを確かめよ。

4.6 条件つき確率と乗法定理

(1)条件つき確率(conditional probability)

ある事象がおこったという条件の下で別の事象のおこる確率

$$Pr(A_2|A_1) = \frac{Pr(A_1A_2)}{Pr(A_1)}, \qquad Pr(A_1|A_2) = \frac{Pr(A_1A_2)}{Pr(A_2)} \quad (4.6)$$

[例題 5]サイコロを順次 2 回投げる実験の場合,

事象A_1：1 回目に 3 の目が出る，事象A_2：出目の和が 7 以上である

とする場合，(4.6)式により条件付き確率を求めよ。

(2)乗法定理：積事象の確率の計算規則

(4.6)式により

$$Pr(A_1A_2) = Pr(A_2|A_1)\,Pr(A_1) = Pr(A_1|A_2)\,Pr(A_2) \quad (4.7)$$

事象A_1と事象A_2が独立の場合,

$Pr(A_2|A_1) = Pr(A_2)\,, Pr(A_1|A_2) = Pr(A_1)$ であるので,

$$Pr(A_1A_2) = Pr(A_1)\,Pr(A_2) \quad (4.8)$$

[例題 6]袋に「赤玉 3 個，白玉 5 個」が入っている。1 個の玉を取り出してその色を確認した後袋のなかに戻さずにもう一度 1 個の玉を取り出す。このとき「1 回目が赤玉で，2 回目が白玉である」確率はいくらか。

4.7 ベイズの定理(Bayes' Theorem)

いま互いに排反する複数の原因$x_i(i=1,2,\cdots,m)$の集まりと，それらのそれぞれから生じ得る互いに排反する複数の結果$y_j(j=1,2,\cdots,n)$の集まりがあるとし，さらに原因x_iのおこる確率$Pr(x_i)$と，原因x_iの下で結果y_jのおこる条件つき確率$Pr(y_j|x_i)$が与えられているとする。

はじめに，原因x_iと結果y_jとが同時におこる確率，すなわち，x_iとy_jの同時確率$Pr(x_iy_j)$は，確率の乗法定理から，

$$Pr(x_iy_j) = Pr(y_j|x_i)\,Pr(x_i) \qquad (4.9)$$

である。ところで，いま「1つの結果がおこった後における，その結果の原因」に関心を持ち，「ある結果y_jがおこったとき，その結果の原因が特定のものx_iである確率$Pr(x_i|y_j)$を求める」ことを考える。この条件つき確率を求めるには，まず(4.9)式の右辺を形式的に書き換えて

$$Pr(x_iy_j) = Pr(x_i|y_j)\,Pr(y_j) \qquad (4.10)$$

とし，これから

$$Pr(x_i|y_j) = \frac{Pr(x_iy_j)}{Pr(y_j)} \qquad (4.11)$$

とすればよい。(4.11)式の右辺の分子$Pr(x_iy_j)$は同時確率であり，これは(4.9)式により求められる。また(4.11)式の右辺の分母$Pr(y_j)$は周辺確率であり，これは次式により求められる。

$$Pr(y_j) = \sum_{k=1}^{m} Pr(y_j|x_k)\,Pr(x_k) \qquad (4.12)$$

したがって，(4.11)式は

$$Pr(x_i|y_j) = \frac{Pr(y_j|x_i)\,Pr(x_i)}{\sum_{k=1}^{m} Pr(y_j|x_k)\,Pr(x_k)} \qquad (4.13)$$

と書くことができる。これがベイズの定理である。

このように，ベイズの定理は，各原因のおこる確率$Pr(x_i)$と，ある原因の下である結果のおこる確率$Pr(y_j|x_i)$が与えられているとき，それらの確率から特定の結果を条件とするある原因の確率$Pr(x_i|y_j)$を求めるためのものである。ベイズの定理では，$Pr(x_i)$を事前確率(prior probability)，$Pr(y_j|x_i)$を尤度(likelihood)，おらび$Pr(x_i|y_j)$を事後確率(posterior probability)という。次の実例を通してベイズの定理に関わる以上の諸確率を具体的に求めて見ることにしよう。

[例題 1]生産管理問題

ある工場では 2 つのグループ(A グループと B グループ)に分かれて 1 つの部品を生産している。両グループはその部品を 1 日同量生産しており，いままでのデータにより把握されている不良率は A グループが 3%, B グループが 5%である。1 つの不良品(あるいは良品)が見つかったときそれが A グループ(あるいは B グループ)の生産したものである確率はいくらであろうか。

①事前確率(原因の発生確率)

この問題で原因と結果は次のように分けられる。

原因(x_i)： x_1(A グループの生産)， x_2(B グループの生産)

結果(y_i)： y_1(不良品)， y_2(良　品)

両グループは 1 日同量を生産するので，事前確率 $Pr(x_1) = Pr(x_2) = 0.5$ となる。

②尤度(ある原因の下である結果のおこる確率)

既知の不良率のデータから，尤度は次のようになる。

$Pr(y_1|x_1) = 0.03, \quad Pr(y_2|x_1) = 0.97$

$Pr(y_1|x_2) = 0.05, \quad Pr(y_2|x_2) = 0.95$

例えば，A グループ(x_1)が不良品(y_1)を生産する確率は 3%であるので，$Pr(y_1|x_1) = 0.03$となる。

③同時確率(ある原因とある結果が同時におこる確率)

同時確率は「事前確率×尤度」として求められる。

$$Pr(x_1y_1) = Pr(y_1|x_1)\,Pr(x_1) = 0.5 \times 0.03 = 0.015$$

$$Pr(x_2y_1) = Pr(y_1|x_2)\,Pr(x_2) = 0.5 \times 0.05 = 0.025$$

$$Pr(x_1y_2) = Pr(y_2|x_1)\,Pr(x_1) = 0.5 \times 0.97 = 0.485$$

$$Pr(x_2y_2) = Pr(y_2|x_2)\,Pr(x_2) = 0.5 \times 0.95 = 0.475$$

④周辺確率(原因と関係なしに特定の結果のおこる確率)

確率の乗法定理により，２つの周辺確率は次のように求められる。

$$Pr(y_1) = Pr(y_1|x_1)\,Pr(x_1) + Pr(y_1|x_2)\,Pr(x_2)$$
$$= Pr(x_1y_1) + Pr(x_2y_1) = 0.015 + 0.025 = 0.04$$

$$Pr(y_2) = Pr(y_2|x_1)\,Pr(x_1) + Pr(y_2|x_2)\,Pr(x_2)$$
$$= Pr(x_1y_2) + Pr(x_2y_2) = 0.485 + 0.475 = 0.96$$

⑤事後確率(特定の結果が判明したとき，それがある原因から由来する確率)

ベイズの定理(4.13)を用いると，次のように事後確率が求められる。

$$Pr(x_1|y_1) = Pr(x_1y_1) \div Pr(y_1) = 0.015 \div 0.04 = 0.375$$

$$Pr(x_2|y_1) = Pr(x_2y_1) \div Pr(y_1) = 0.025 \div 0.04 = 0.625$$

$$Pr(x_1|y_2) = Pr(x_1y_2) \div Pr(y_2) = 0.485 \div 0.96 = 0.505$$

$$Pr(x_2|y_2) = Pr(x_2y_2) \div Pr(y_2) = 0.475 \div 0.96 = 0.495$$

これにより，例えば，１つの不良品(y_1)が見つかったとき，それがＡグループ(x_1)の生産したものである確率 $\Pr(x_1|y_1) = 0.375$ であることがわかる。

[諸確率の簡便計算表]

ベイズの定理(4.13)式を用いれば事後確率を求めることができるが，その計算プロセスは少々煩雑である。そこで，簡便計算表を利用すれば極めて容易にベイズの定理に関わる諸確率が求められる。

表 4.1：簡便計算表

原因(X)						
結果(Y)						

table1	結果→	y1	y2	y3	y4	y5
原因↓	事前確率	尤　　度				
x1						
x2						
x3						
x4						
x5						

table2	結果→	y1	y2	y3	y4	y5
原因↓		同時確率				
x1						
x2						
x3						
x4						
x5						
周辺確率						

table3	結果→	y1	y2	y3	y4	y5
原因↓		事後確率				
x1						
x2						
x3						
x4						
x5						
合　計						

簡便計算表の中で求められる諸確率を簡潔に表すために，次のような記号を用いることにする。

$$\left.\begin{array}{ll} \text{事前確率} & Pr(x_i) = p_i \\ \text{尤　　度} & Pr(y_j|x_i) = \lambda_{ij} \\ \text{同時確率} & Pr(x_i y_j) = \mu_{ij} \\ \text{周辺確率} & Pr(y_j) = q_j \\ \text{事後確率} & Pr(x_i|y_j) = \tau_{ij} \end{array}\right\} \quad (4.14)$$

事前確率(p_i)と尤度(λ_{ij})は既知であるので，この２つの確率を用いると，下の(4.15)式により同時確率(μ_{ij})，周辺確率(q_j)，および事後確率(τ_{ij})が順々に求められる。

$$\left.\begin{array}{ll} \text{同時確率} & \mu_{ij} = p_i \times \lambda_{ij} \\ \text{周辺確率} & q_j = \sum \mu_{ij} \\ \text{事後確率} & \tau_{ij} = \mu_{ij} \div q_j \end{array}\right\} \quad (4.15)$$

簡便計算表を用いて前掲の[生産管理問題]に関わる諸確率を求めてみると，次のような数値が得られる。

表 4.2：生産管理問題の簡便計算表

原因(X)	x1(Aグループ)	x2(Bグループ)			
結果(Y)	y1(不良品)	y2(良品)			

table1	結果→	y1	y2	y3	y4	y5
原因↓	事前確率	尤　　度				
x1	0.5	0.03	0.97			
x2	0.5	0.05	0.95			
x3						
x4						
x5						

table2	結果→	y1	y2	y3	y4	y5
原因↓		同時確率				
x1		0.015	0.485			
x2		0.025	0.475			
x3						
x4						
x5						
周辺確率		0.040	0.960			

table3	結果→	y1	y2	y3	y4	y5
原因↓		事後確率				
x1		0.375	0.505			
x2		0.625	0.495			
x3						
x4						
x5						
合　計		1.000	1.000			

[練習問題 1]

ある電子部品で競合するメーカー3 社の市場シェアは，P 社 50%，Q 社 30%，R 社 20%である。この電子部品の製造不良率は，P 社 10%，Q 社 20%，R 社 25%と知られている。今，どのメーカー製のものかわからずに市場でこの電子部品を 1 個購入したところ，それが不良品であった場合，それがP 社製，Q 社製，R 社製である確率はそれぞれいくらであるか。

[練習問題 2]

ある県のシートベルト着用率は 82%である。この県の今年上半期の交通事故に関するデータによると，シートベルト着用者の場合「死亡率 12%，重症率 25%，軽傷率 63%」であるのに対して，シートベルト非着用者の場合は「死亡率 32%，重症率 53%，軽傷率 15%」であった。Excel を用いて簡便計算表を作成しベイズの定理に関わる諸確率を求めよ。

5 章 確率変数と確率分布

5.1 確率変数(random variable)

各標本点に対応してその値が決まるような変数を確率変数という。

例えば，2 枚の硬貨を投げるとき，「裏が出たら 0，表が出たら 1」とし，その合計を x とすると，標本点および x の確率は次のようになる。

標本点	裏・裏	裏・表	表・裏	表・表
$\boldsymbol{x}$	0	1	1	2
確率	0.25	0.25	0.25	0.25

上の表は確率変数値と確率との対応を明示してくれるが，普通は次のような実用的な表記を用いる。

$$Pr(x = 0) = 0.25,\ Pr(x = 1) = 0.5, Pr(x = 2) = 0.25$$

上式の中の Pr は確率を意味する記号である。上式は「$x = 0$となる確率は 0.25」，「$x = 1$となる確率は 0.5」，「$x = 2$となる確率は 0.25」であることを表す。

・2 種類の確率変数

① 離散確率変数(discrete probability distribution)：

変数値がとびとびの値である場合の確率変数(デジタル変数)

例えば，サイコロなげの場合の出目の数字，一定期間内の事故件数，等々

② 連続確率変数(continuous probability distribution)：

変数値が連続値と考えられる場合の確率変数(アナログ変数)

例えば，音声，アナログ電波，等々

5.2 確率分布(probability distribution)

・各変数値に対して確率が対応する仕方(ルール)

・2 種類の確率分布

① 離散確率分布：2 項分布，ポワソン分布，超幾何分布，等々

② 連続確率分布：三角分布，正規分布，指数分布，等々

(1) 2 項分布(binomial distribution)

* 状況の記述

① 標本空間に 2 事象(A or B)のみが存在し,

② A 事象の発生確率がp (B 事象の発生確率は$1-p$)である実験をn回実施したとき,

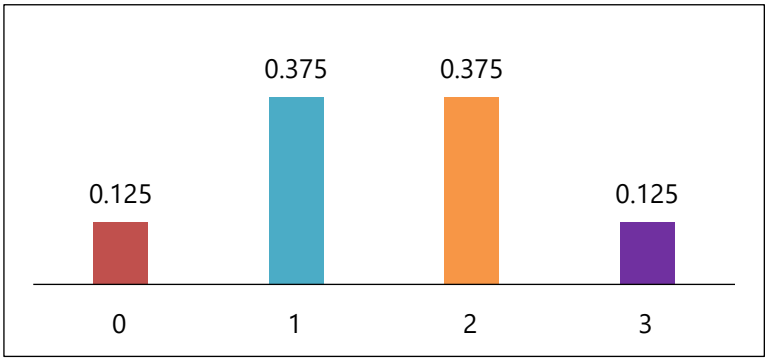
③ A 事象がx回おこる確率の分布を 2 項分布という。

* 計算式

$$\underset{b}{Pr}(x,n,p) = {}_nC_x\ p^x(1-p)^{n-x} = \frac{n!}{x!\,(n-x)!}p^x(1-p)^{n-x} \qquad (5.1)$$

例えば, 3 回硬貨を投げて「表の出る回数」を x とすると, ここには 2 つの事象[表が出る(A) or 裏が出る(B)]から成る標本空間が存在し, 表の出る確率が 0.5(裏の出る確率 0.5)であるので, x (=0, 1, 2, 3)回表の出る確率は次のように求められる。

$$\underset{b}{Pr}(x=0,\ n=3,\ p=0.5) = {}_3C_0\ 0.5^0\ (1-0.5)^{3-0} = 0.125$$
$$\underset{b}{Pr}(x=1,\ n=3,\ p=0.5) = {}_3C_1\ 0.5^1\ (1-0.5)^{3-1} = 0.375$$
$$\underset{b}{Pr}(x=2,\ n=3,\ p=0.5) = {}_3C_2\ 0.5^2\ (1-0.5)^{3-2} = 0.375$$
$$\underset{b}{Pr}(x=3,\ n=3,\ p=0.5) = {}_3C_3\ 0.5^3\ (1-0.5)^{3-3} = 0.125$$



[例題 1] 2 項分布の確率を求めるプログラムを作成せよ。

(2)ポワソン分布(Poisson distribution)

* 大量の観察回数の下で極めて稀におこる事象の確率分布

例えば，１日当りの飛行機墜落件数，短時間における自然放射線の計数値，等々

* 2 項分布において，平均発生件数(m)で，$p \to 0(\therefore n \to \infty)$とした場合の確率分布

* 計算式

$$\underset{p}{Pr}(x,m) = \frac{m^x\, e^{-m}}{x!} \qquad (5.2)$$

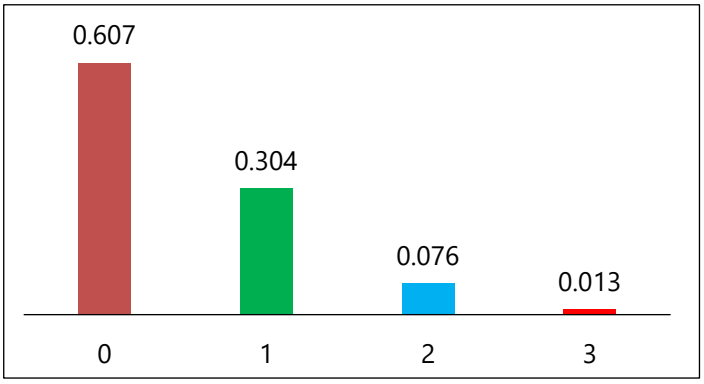
例えば，ある県の交通事故で，１日の平均死亡事故件数(m)が 0.5 件である場合，この県で１日の死亡事故件数$(x = 0,1,2,3)$の確率は次のように求められる。

$$\underset{p}{Pr}(x = 0, m = 0.5) = \frac{0.5^0\, e^{-0.5}}{0!} = 0.607$$

$$\underset{p}{Pr}(x = 1, m = 0.5) = \frac{0.5^1\, e^{-0.5}}{1!} = 0.304$$

$$\underset{p}{Pr}(x = 2, m = 0.5) = \frac{0.5^2\, e^{-0.5}}{2!} = 0.076$$

$$\underset{p}{Pr}(x = 3, m = 0.5) = \frac{0.5^3\, e^{-0.5}}{3!} = 0.013$$



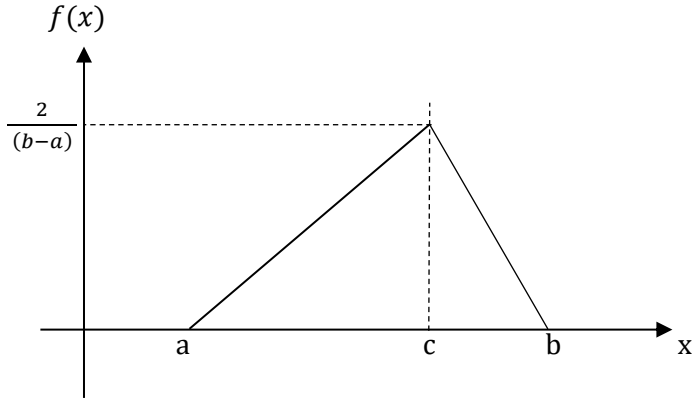
[例題 2]ポワソン分布の確率を求めるプログラムを作成せよ。

(3)三角分布(triangular distribution)

採取可能なサンプルデータの数に限りのある母集団を表すために用いられるのが三角分布である。例えば，ある建設工事の完成所要期間について，「最短値(a)，最長値(b)，最頻値（c：ピーク値)」として推定することができる場合の連続確率分布が三角分布である。この分布は「最短値から最頻値まで直線的に増加し，最頻値から最長値まで直線的に減少する(三角形の形状となる)」ので三角分布と呼ばれる。

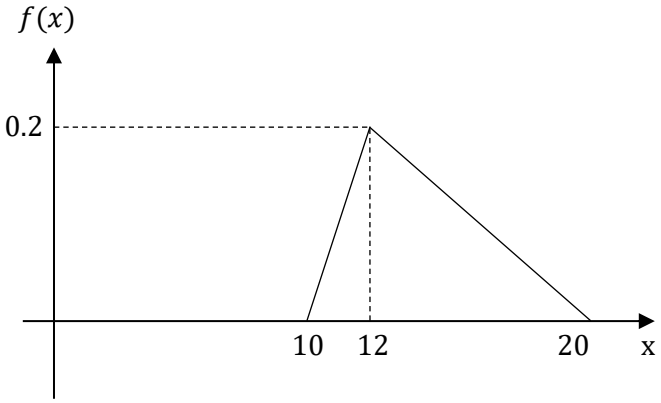
三角分布の確率密度関数は以下のように表される。

$$f(x)=\begin{cases} \dfrac{2(x-a)}{(b-a)(c-a)} & for\ a\le x<c \\ \dfrac{2}{(b-a)} & for\ x=c \\ \dfrac{2(b-x)}{(b-a)(b-c)} & for\ c<x\le b \end{cases} \qquad (5.3)$$



したがって，ある建設工事の完成所要期間が「最短 10 日(a=10)，最長 20 日(b=20)，最頻 12 日(c=12)」の場合の三角分布の確率密度関数は以下のようになる。

$$f(x) = \left\{ \begin{array}{cc} 0.1(x-10) & for\ 10 \le x < 12 \\ 0.2 & for\ x = 12 \\ 0.025(20-x) & for\ 12 < x \le 20 \end{array} \right\}$$



三角分布は連続確率分布であるので特定の変数値の確率は求められず，2 つの変数値の間における三角形の下の面積が確率となる。例えば，建設工事の完成に 10 以上 12 日以下かかる確率は

$$Pr(10 \le x \le 12) = \frac{(12-10) \times 0.2}{2} = 0.2$$

となる。また，建設工事の完成に 12 日以上 20 日以下かかる確率は

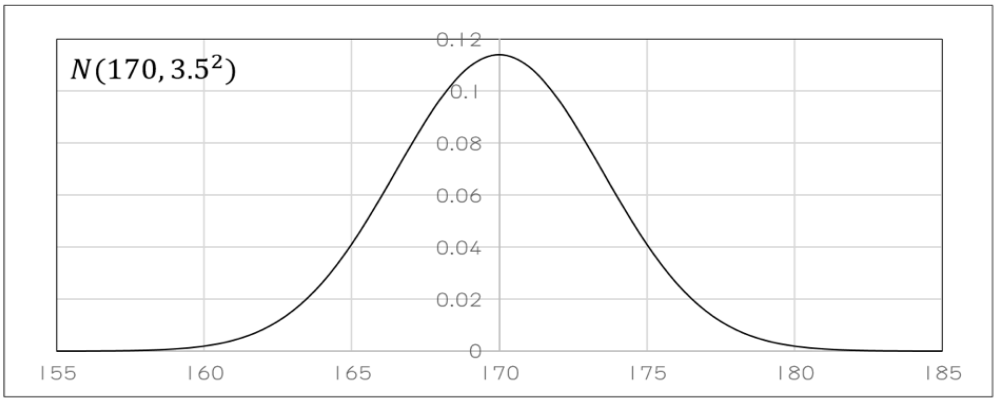
$$Pr(12 \le x \le 20) = \frac{(20-12) \times 0.2}{2} = 0.8$$

となる。

[例題 3]三角分布の確率を求めるプログラムを作成せよ。

(4)正規分布(Normal distribution)

標本の数が多い$(n \geq 25)$場合，大数の法則によりその標本の分布は正規分布となる。例えば，東京都内の高校３年生 500 人の身長を測り，その平均値と標準偏差を求めて見たところ，平均値が 170cm，標準偏差が 3.5cm であった場合，その分布は正規分布となり，$N(170, 3.5^2)$と表わせる。



＊ 特徴

・「平均値(μ)と標準偏差(σ)」の 2 つのパラメータによって正規分布の形が完全に決まる。したがって，$N(\mu, \sigma^2)$と表記する。

・正規分布は平均値を中心に左右対称の形をとる。

・正規分布をする代表的な数値に偏差値(平均値=50，標準偏差=10)がある。

＊ 確率密度関数(probability density function)

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (5.4)$$

＊ 正規分布の確率

$$Pr(x < A) = \int_{-\infty}^{A} f(x)\,dx = \int_{-\infty}^{A} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (5.5)$$

正規分布の確率は，確率密度関数を確率変数の任意の区間について積分することによって求められる。

* 標準化(standadization)

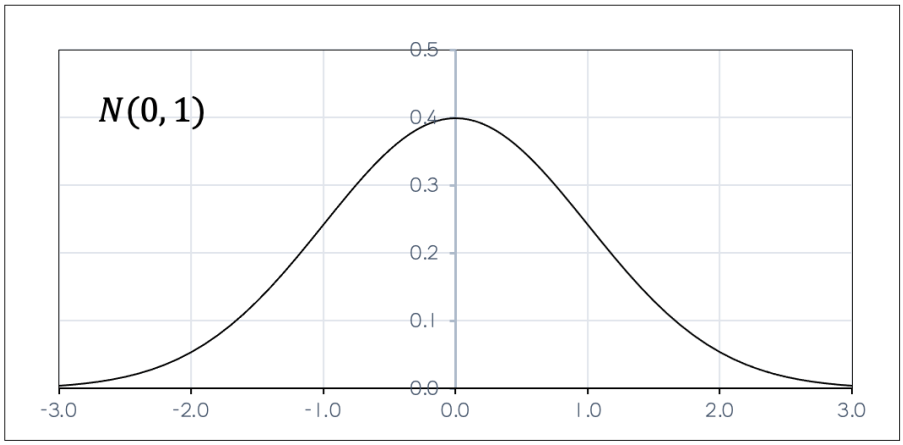
元の変数(x) ----------(標準化)----------> 標準化変量(z)

$N(\mu, \sigma^2)$ $\qquad$ $N(0,1)$

$$z = \frac{x-\mu}{\sigma} \qquad (5.6)$$

* 標準正規分布 : $N(0,1)$

正規分布にしたがう確率変数xを(5.6)式を用いて線形変換すると，標準化変量(z)は標準正規分布にしたがう確率変数となる。



正規分布に従う確率変数(x)が任意の区間$(a \le x \le b)$にある確率$Pr(a \le x \le b)$を求めるためには，(5.6)式によりxをzに変換して，

$$Pr(a \le x \le b) = Pr\left\{\frac{a-\mu}{\sigma} \le z \le \frac{b-\mu}{\sigma}\right\} \qquad (5.7)$$

を用いればよい。

例えば，$N(20, 2^2)$において $Pr(21 \le x \le 23)$を求めるには，まず(5.7)のように変数値を標準化する。

$$Pr(21 \le x \le 23) = Pr\left\{\frac{21-20}{2} \le z \le \frac{23-20}{2}\right\} = \Pr(0.5 \le z \le 1.5)$$

ここで巻末の標準正規分布表を用いるためには

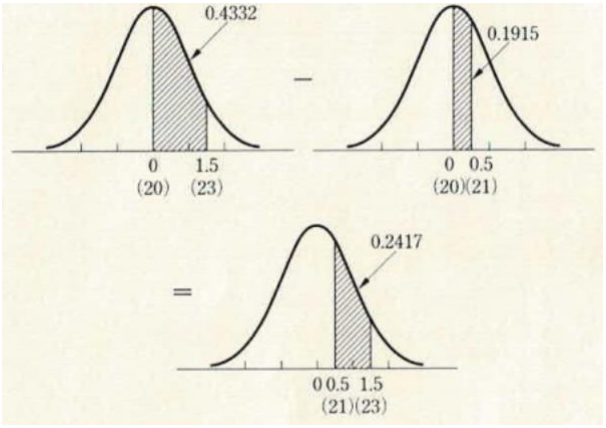
$$Pr(0.5 \leq z \leq 1.5) = Pr(0.0 \leq z \leq 1.5) - Pr(0.0 \leq z \leq 0.5)$$

と計算すればよい。ここで標準正規分布表より

$$Pr(0.0 \leq z \leq 1.5) = 0.4332, \qquad Pr(0.0 \leq z \leq 0.5) = 0.1915$$

がわかる。よって,

$$Pr(21 \leq x \leq 23) = 0.4332 - 0.1915 = 0.2417$$

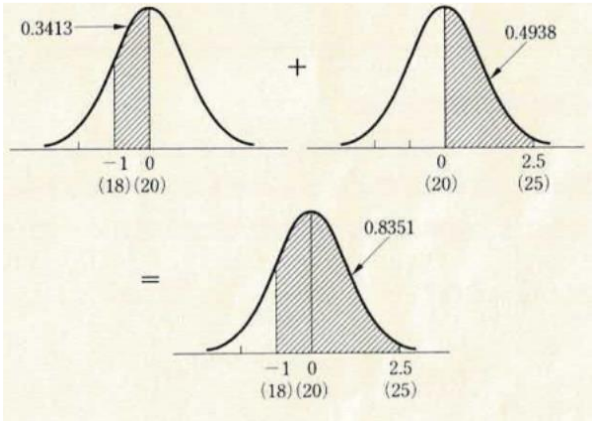


次に，$Pr(18 \leq x \leq 25)$を求める問題を考える。

$$Pr(18 \leq x \leq 25) = Pr\left\{\frac{18-20}{2} \leq z \leq \frac{25-20}{2}\right\} = \Pr(-1.0 \leq z \leq 2.5)$$

であり, 左右対称性から$\Pr(-1.0 \leq z \leq 0.0) = \Pr(0.0 \leq z \leq 1.0)$であることに注意して, 標準正規分布表により,

$$Pr(18 \leq x \leq 25) = 0.3413 + 0.4938 = 0.8351$$



[例題 4]正規分布の確率を求めるプログラムを作成せよ。

5.3 期待値と分散

(1)期待値(expected value)

確率変数の平均値を期待値といい，「変数値と確率との積の和」により求められる。

離散確率変数の場合，$\mu = E(x) = \sum_x x\,Pr(x)$　(5.8)

連続確率変数の場合，$\mu = E(x) = \int_{-\infty}^{+\infty} x\,f(x)\,dx$　(5.9)

(2)分散(variance)

確率分布のバラツキの度合いを表す指標であり，「偏差(データと平均値との差)の 2 乗値の期待値」として求められる。

$$\sigma^2 = E[(x-\mu)^2] \quad (5.10)$$

離散確率変数の場合，$\sigma^2 = V(x) = \sum_x (x-\mu)^2\,Pr(x)$　(5.11)

連続確率変数の場合，$\sigma^2 = V(x) = \int_{-\infty}^{+\infty} (x-\mu)^2\,f(x)\,dx$　(5.12)

(5.10)式を展開し整理すると，次のような「分散の簡便計算式」が得られる。

$$\sigma^2 = V(x) = E[(x-\mu)^2] = \cdots\cdots = E(x^2) - \mu^2 \quad (5.13)$$

[練習問題 1]

サイコロ 2 個を同時に投げて出る目の和を 4 で割ったときの余りを x する。

(1) x の確率分布を求めよ。

(2) x の分布の期待値と分散を求めよ。

[練習問題 2]

壺の中に 8 個の赤い球と 4 個の白い球が入っている。この壺の中から同時に 2 個の球を無作為に抜き取るものとするとき，次の問に答えよ。

(1) 抜き取られた赤い球の数が 0, 1, 2 である確率を求めよ。

(2) 赤い球 1 個当たり 1000 円，白い球 1 個当たり 500 円をもらえるとすれば，もらえる金額の期待値はいくらか。また分散はいくらか。

6 章 標本分布

6.1 母集団と標本

いま，ある TV 局で特定の番組に対する視聴率を調べようとする場合，TV 局で調べたいと思うのは，全視聴者の視聴率である。ところで，膨大な数の視聴者の全部について調べることは，事実上不可能である。そこで TV 局は，視聴者全体の中から何十人あるいは何百人の人たちを選んで，それらの人たちについて特定の番組を視聴しているか否かを調べ，それに基づいて視聴者全体では視聴率はどうであろうかを推測するのである。

この例の場合，TV 局が知りたいと思っている視聴者全体の集まり，一般には推測の対象となる全体の集団のことを「母集団(population)」といい，これに対して母集団の中から選ばれる一部分の集まりのことを「標本(sample)」という。そして標本についての調査，すなわち「標本調査(sample survey)」に基づいて母集団の性質を推測することを「統計的推論(statistical inference)」という。

母集団全体を調べずに標本調査に頼らなければならない場合は極めて多く，その主な理由としては次の３つを挙げることができる。

第 1 に，母集団が実在していないということがある。例えば，病気にたいする新薬の効果を知ろうとする場合，母集団はその病気にかかった人全体の集団であり，その中にはまだ現実にその病気にかかっていない人も含まれるのである。

第 2 に，調査することは破壊を意味することがある。例えば，これから生産する予定の新車の強度を調べる場合，現実に破壊テストが行われており，母集団全体を破壊することは考えられない。

第 3 に，母集団全体を調べることが物理的に可能であっても，時間や資源の制約からそれができない場合がある。マスコミによる世論調査のような場合がその典型例である。

このような理由から，一部分の標本から母集団全体に適用できるような情報を引き出すべく標本調査が行われるのである。

ところで，標本調査においては，母集団から選び出される標本が母集団を偏らずに正しく代表するものでなければならない。そのような代表的標本を選ぶための方法はいろいろあるが，もっとも単純で分かりやすい方法は「単純無作為抽出(simple random sampling)」である。

・乱数表：乱数サイ(正 20 面体のサイコロ)を振って出た数字を並べた数表

28 89 65 87 08	13 50 63 04 23	25 47 57 91 13	52 62 24 19 94	91 67 48 57 10
30 29 43 65 42	78 66 28 55 80	47 46 41 90 08	55 98 78 10 70	49 92 05 12 07
95 74 62 60 53	51 57 32 22 27	12 72 72 27 77	44 67 32 23 13	67 95 07 76 30
01 85 54 96 72	66 86 65 64 60	56 59 75 36 75	46 44 33 63 71	54 50 06 44 75
10 91 46 96 86	19 83 52 47 53	65 00 51 93 51	30 80 05 19 29	56 23 27 19 03
05 33 18 08 51	51 78 57 26 17	34 87 96 23 95	89 99 93 39 79	11 28 94 15 52
04 43 13 37 00	79 68 96 26 60	70 39 83 66 56	62 03 55 86 57	77 55 33 62 02
05 85 40 25 24	73 52 93 70 50	48 21 47 74 63	17 27 27 51 26	35 96 29 00 45
84 90 90 65 77	63 99 25 69 02	09 04 03 35 78	19 79 95 07 21	02 84 48 51 97
28 55 53 09 48	86 28 30 02 35	71 30 32 06 47	93 74 21 86 33	49 90 21 69 74
89 83 40 69 80	97 96 47 59 97	56 33 24 87 36	17 18 16 90 46	75 27 28 52 13
73 20 96 05 68	93 41 69 96 07	97 50 81 79 59	42 37 13 81 83	92 42 85 04 31
10 89 07 76 21	40 24 74 36 42	40 33 04 46 24	35 63 02 31 61	34 59 43 36 96
91 50 27 78 37	06 06 16 25 98	17 78 80 36 85	26 41 77 63 37	71 63 94 94 33
03 45 44 66 88	97 81 26 03 89	39 46 67 21 17	98 10 39 33 15	61 63 00 25 92
89 41 58 91 63	65 99 59 97 84	90 14 79 61 55	56 16 88 87 60	32 15 99 67 43
13 43 00 97 26	16 91 21 32 41	60 22 66 72 17	31 85 33 69 07	68 49 20 43 29
71 71 00 51 72	62 03 89 26 32	35 27 99 18 25	78 12 03 09 70	50 93 19 35 56
19 28 15 00 41	92 27 73 40 38	37 11 05 75 16	98 81 99 37 29	92 20 32 39 67
56 38 30 92 30	45 51 94 69 04	00 84 14 36 37	95 66 39 01 09	21 68 40 95 79
39 27 52 89 11	00 81 06 28 48	12 08 05 75 26	03 35 63 05 77	13 81 20 67 58
73 13 28 58 01	05 06 42 24 07	60 60 29 99 93	72 93 78 04 36	25 76 01 54 03
81 60 84 51 57	12 68 46 55 89	60 09 71 87 89	70 81 10 95 91	83 79 68 20 66
05 62 98 07 85	07 79 26 69 61	67 85 72 37 41	85 79 76 48 23	61 58 87 08 05
62 97 16 29 18	52 16 16 23 56	62 95 80 97 63	32 25 34 03 36	48 84 60 37 65

[乱数表の使い方]

500 人の母集団からランタムに 10 人を抽出する。

① 各個人に背番号(000～499)を付与する。

② 乱数表から 3 桁の乱数を抜き取る。

乱数表のどこから乱数を拾っても構わないので，ここでは乱数表の左上から下に進みながら 3 桁の乱数を拾うことにする。ただ，500 以上の乱数が選ばれた場合はその乱数は捨てる。

288　302　~~957~~　018　109　053　044　058　~~849~~　285　~~898~~

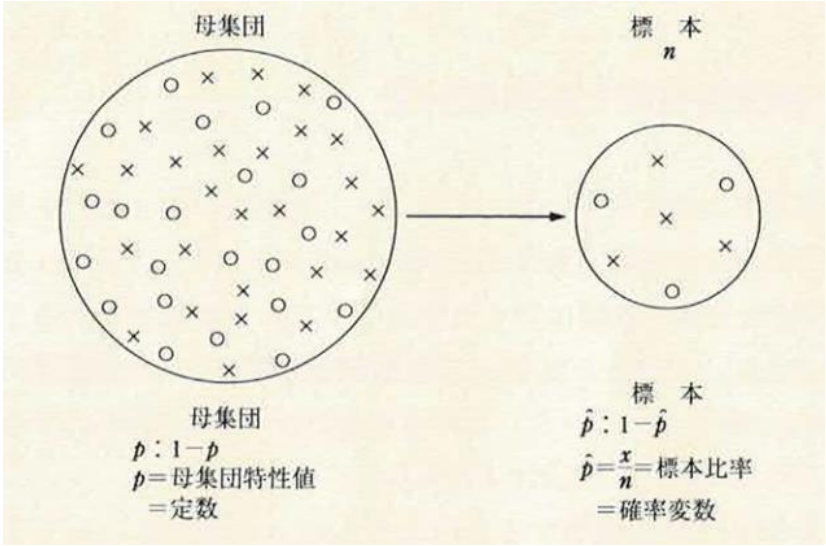
~~732~~　108　~~915~~　034 ⇒ この背番号の人を標本とする。

[例題 1]

1 桁の乱数 2 個の平均をとる実験を 10 回繰り返して，その平均値と標準偏差を求めよ。

理論値は，平均が 4.50，標準偏差が 2.03 である。

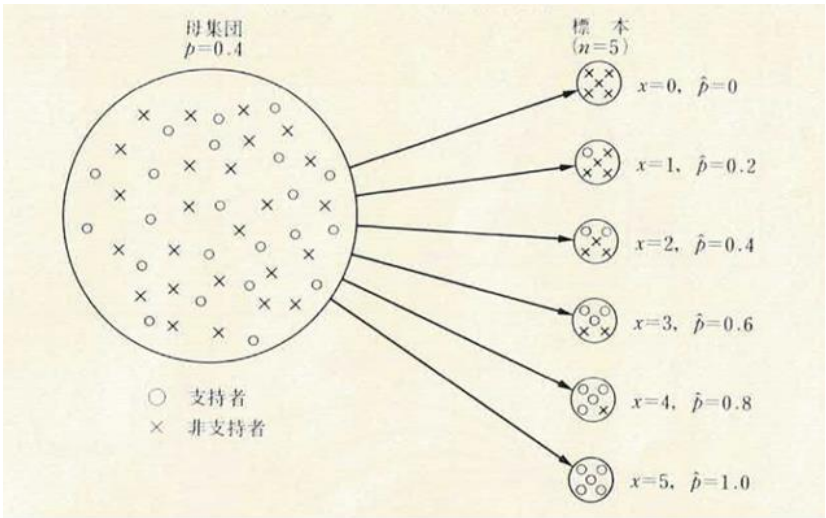
6.2　母集団特性値と標本統計量



・標本分布：標本統計量の確率分布

例えば，母集団の内閣支持率を $p = 0.4$と仮定し，

大きさ $n = 5$の標本から調べた内閣支持率$(\hat{p})$の分布



・$\hat{p}$ (標本比率)の標本分布

標本のなかの支持者の数(x)は $n = 5, p = 0.4$ の２項分布をする確率変数である。

$\therefore Pr(x, n = 5, p = 0.4) = {}_5C_x \, 0.4^x \, (1 - 0.4)^{5-x} = Pr(\hat{p})$

$$Pr(x=0,n=5,p=0.4)=\Pr(\hat{p}=0.0)=0.0778$$
$$Pr(x=1,n=5,p=0.4)=\Pr(\hat{p}=0.2)=0.2592$$
$$Pr(x=2,n=5,p=0.4)=\Pr(\hat{p}=0.4)=0.3456$$
$$Pr(x=3,n=5,p=0.4)=\Pr(\hat{p}=0.6)=0.2304$$
$$Pr(x=4,n=5,p=0.4)=\Pr(\hat{p}=0.8)=0.0768$$
$$Pr(x=5,n=5,p=0.4)=\Pr(\hat{p}=1.0)=0.0102$$

表 7.2 $\hat{p}$ の標本分布

$\hat{p}$	0.0	0.2	0.4	0.6	0.8	1.0
$\Pr(\hat{p})$	0.0778	0.2592	0.3456	0.2304	0.0768	0.0102

6.3　標本比率$(\hat{p})$の標本分布

① 標本のなかの支持者の数(x)は 2 項分布に従う確率変数である。

$$\therefore E(x)=np, V(x)=np(1-p)=npq \qquad (6.1)$$

② 標本の数が大きければ，xは正規分布 $N(np,npq)$ にしたがう確率変数となる。

$$\therefore E\left(\frac{x}{n}\right)=E(\hat{p})=\left(\frac{1}{n}\right)E(x)=\left(\frac{1}{n}\right)np=p \qquad (6.2)$$

$$V\left(\frac{x}{n}\right)=V(\hat{p})=\left(\frac{1}{n^2}\right)V(x)=\left(\frac{1}{n^2}\right)npq=\frac{pq}{n} \qquad (6.3)$$

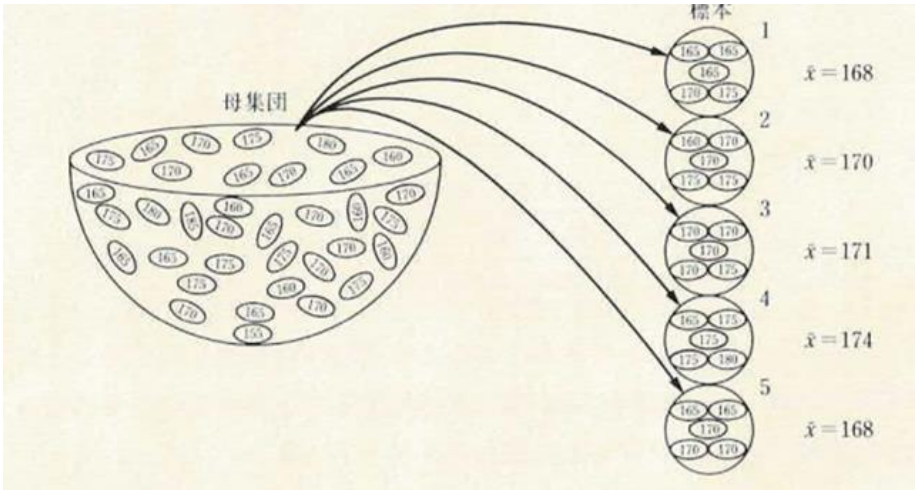
$\hat{p}$の標本分布は，標本の数が大きければ，近似的に$N\left(p,\frac{pq}{n}\right)$にしたがうことになる。

$\therefore z=\frac{\hat{p}-p}{\sqrt{\frac{pq}{n}}}$は，近似的に$N(0,1)$にしたがう確率変数となる。

[例題 2]

今年の勝率 6 割の投手が来年 20 試合に登板して 15 勝以上する確率を求めよ。

6.4 標本平均値$(\bar{x})$の標本分布－平均値と分散



$$E(\bar{x}) = E\left[\frac{x_1 + x_2 + \cdots + x_n}{n}\right] = \frac{1}{n}E(x_1 + x_2 + \cdots + x_n)$$

$$= \frac{1}{n}[E(x_1) + E(x_2) + \cdots + E(x_n)] = \frac{1}{n}n\mu = \mu \qquad (6.4)$$

$$V(\bar{x}) = V\left[\frac{x_1 + x_2 + \cdots + x_n}{n}\right] = \frac{1}{n^2}V(x_1 + x_2 + \cdots + x_n)$$

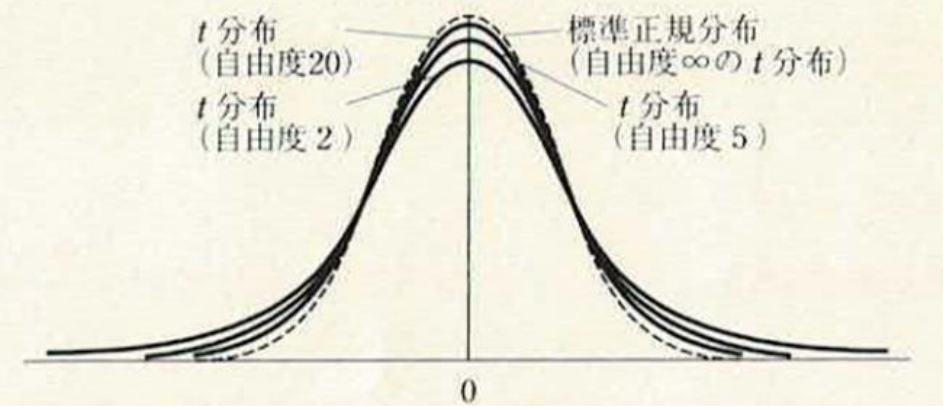
$$= \frac{1}{n^2}[V(x_1) + V(x_2) + \cdots + V(x_n)] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n} \qquad (6.5)$$

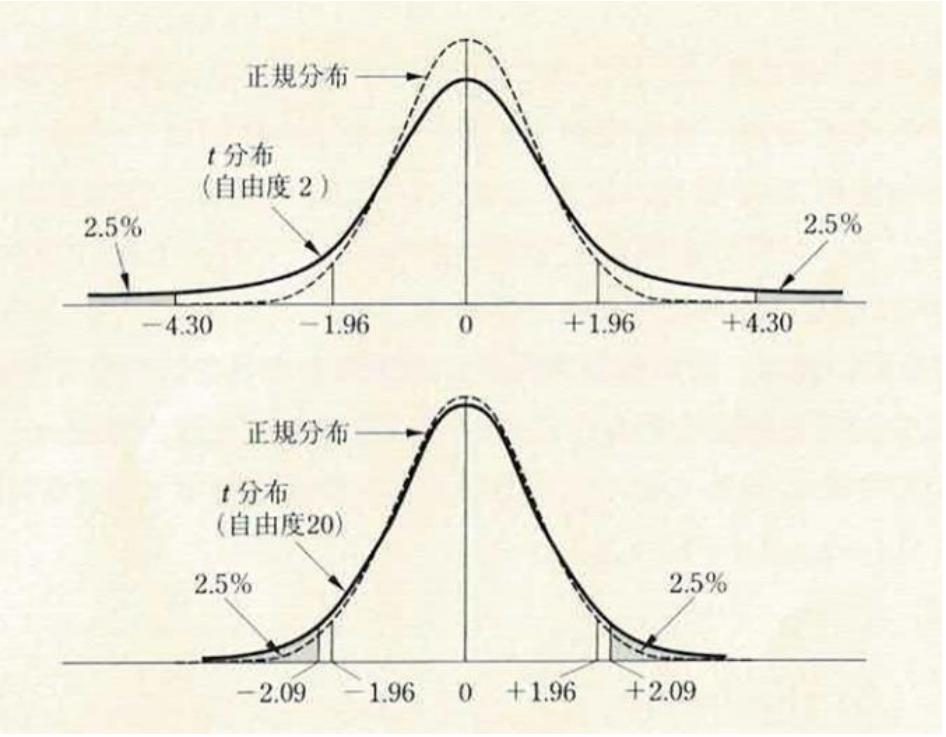
$\bar{x}$ の標本分布は，標本の数が大きければ， $N(\mu, \frac{\sigma^2}{n})$ にしたがうことになる。 ∴ $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ は，$N(0,1)$にしたがう確率変数となる。

[例題 3]

母集団の分布が$N(50, 8^2)$であるとき，100 の確率標本で平均値$(\bar{x})$が 49 以下となる確率はいくらか。また，400 の確率標本の場合はどうか。

6.5 t分布(t distribution)





・標本の数(n)が大きければ，$\bar{x}$の標準化変量 $z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ の分布は $N(0,1)$ にしたがう。

$$\therefore Pr\left(-1.96<\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}<1.96\right)=0.95 \qquad (6.6)$$

$$Pr\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}}<\mu<\bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right)=0.95 \qquad (6.7)$$

$$Pr(\mu_1<\mu<\mu_2)=0.95 \qquad (6.8)$$

・標本標準偏差$(\hat{\sigma})$ ：

$$\hat{\sigma}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad (6.9)$$

・σ が未知の場合，$\bar{x}$の標準化変量 $t=\frac{\bar{x}-\mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$ は自由度$(n-1)$の t 分布に従う。

・t 分布の性質

①平均値を中心として左右対称である。

②正規分布に似ているが，正規分布より分布の頂点が低く，裾野が長い。

③ 標本の数(n)が決まれば，分布の形が決まる。

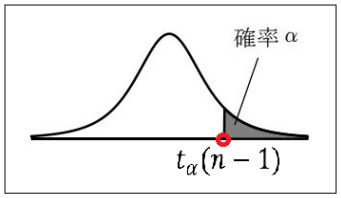
自由度(degree of freedom) =$(n-1)$は t 分布の唯一のパラメータである。

④ 標本の数(n)が小さければ，分布の頂点が段々低くなる。

⑤ 標本の数(n)が無限大になると，t 分布は正規分布に一致する。

・t 値と確率

t 値の表記： $t_\alpha(n-1)$



例えば，　$n=6$，$\alpha=0.025$ の場合，$t_{0.025}(5)=2.571$

$n=6$，$\alpha=0.005$ の場合，$t_{0.005}(5)=4.032$

・t分布の利用

t分布を用いれば，母集団標準偏差σの値がわからない場合でも母集団平均値μについての推論を行なうことができる。

t分布においては，t値が$[-t_{0.05}(n-1), t_{0.05}(n-1)]$の区間内に含まれる確率は90%である。したがって，

$$Pr\{-t_{0.05}(n-1) < t < t_{0.05}(n-1)\} = 0.9 \qquad (6.9)$$

$$Pr\left\{-t_{0.05}(n-1) < \frac{\bar{x}-\mu}{\frac{\hat{\sigma}}{\sqrt{n}}} < t_{0.05}(n-1)\right\} = 0.90 \qquad (6.10)$$

$$Pr\left\{\bar{x} - t_{0.05}(n-1)\frac{\hat{\sigma}}{\sqrt{n}} < \mu < \bar{x} + t_{0.05}(n-1)\frac{\hat{\sigma}}{\sqrt{n}}\right\} = 0.90 \qquad (6.11)$$

となる。同様にして

$$Pr\left\{\bar{x} - t_{0.025}(n-1)\frac{\hat{\sigma}}{\sqrt{n}} < \mu < \bar{x} + t_{0.025}(n-1)\frac{\hat{\sigma}}{\sqrt{n}}\right\} = 0.95 \qquad (6.12)$$

$$Pr\left\{\bar{x} - t_{0.005}(n-1)\frac{\hat{\sigma}}{\sqrt{n}} < \mu < \bar{x} + t_{0.005}(n-1)\frac{\hat{\sigma}}{\sqrt{n}}\right\} = 0.99 \qquad (6.13)$$

となる。

[例題 4]

ある母集団から抽出した9個の標本(28, 42, 39, 62, 88, 52, 55, 21, 72)から母集団平均値を区間推定せよ。

[練習問題 1]

3 割の打撃力をもつ野球選手が 20 打数で2割 5 分以下の打率しかあげられない確率はいくらか。また，100 打数でならばどうか。

[練習問題 2]

長年にわたって収集された統計によると，ある市場で売買された食肉牛の体重は平均 860kg，標準偏差 180kg の正規分布でよく近似できることが分かっている。
これら食肉牛 400 頭からなる無作為標本の平均体重か次のようになる確率はいくらか。

(a)850kg 以下　　(b)845kg 以上 870kg 以下

[練習問題 3]

A 君は通学に，電車とバスを利用する。今までの経験から，電車，乗り換え，バスに要する時間は，平均と標準偏差がそれぞれ以下のような値をとる正規分布にしたがうことがわかっている。

	平均	標準偏差
電　　車	30 分	1 分
乗り換え	5 分	1 分
バ　　ス	15 分	3 分
徒　　歩	5 分	0.5 分

学校が 8 時半に始まるとして，7 時半に電車に乗るとすると，A 君が遅刻する確率はいくらか。ただし各乗り物などの所要時間は互いに関係ないものとする。

7 章 推定と検定

7.1 推定(estimation)

推定は，母集団の特徴を表わす何らかの特性値(parameter)を標本の観察に基づいて推測することてある。例えば，東京都の全有権者(母集団)の中で，内閣を支持している人の割合(特性値)がどれくらいかを，何百人かの人(標本)を選んで調べた結果に基づいて推測することてある。

このような推定において基本的に重要な役割を果たすのが，前章で説明した統計量，およびその分布，すなわち標本分布である。

(1)比率の区間推定

・内閣支持率，TV 番組の視聴率，生産工程の不良率，等々の比率に対する推定

・推定の手順(内閣支持率の推定の場合)

① 標本の数 n ，標本の中の支持者数 x → $\hat{p}=\frac{x}{n}$ (標本比率)の分布は「2 項分布」

② 標本の数が大きければ，$\hat{p}=\frac{x}{n}$ は $N(p,\frac{pq}{n})$ にしたがうことになる。

③ $\hat{p}=\frac{x}{n}$ の標準化変量 $z=\frac{\frac{x}{n}-p}{\sqrt{\frac{pq}{n}}}$ は $N(0,1)$ にしたがう確率変数である。

④ 標準正規分布では

$$Pr\left(-1.64\le\frac{\frac{x}{n}-p}{\sqrt{\frac{pq}{n}}}\le 1.64\right)=0.90 \qquad (7.1)$$

$$Pr\left\{\left(\frac{\frac{x}{n}-p}{\sqrt{\frac{pq}{n}}}\right)^2\le 1.64^2\right\}=0.90 \rightarrow Pr\left\{\left(\frac{x}{n}-p\right)^2\le 1.64^2\frac{pq}{n}\right\}=0.90 \qquad (7.2)$$

$$Pr\{(n+2.69)p^2-(2x+2.69)p+x^2/n\le 0\}=0.90 \qquad (7.3)$$

$$Pr(p_1\le p\le p_2)=0.90 \qquad (7.4)$$

“母集団の比率 p は 90%の確率で p_1 と p_2 の間の値である。”

・推定の用語

区間推定(interval estimation) → 信頼区間(confidence interval)を推定する。

点推定(point estimation) → 信頼値(confidence value)を推定する。

信頼係数(confidence coefficient) → 99%, 95%, 90% → 目的に応じて選択する。

[例題 1]

ある発案に対して，標本として選ばれた 50 人のうち 20 人が賛成した。母集団の比率を区間推定せよ。

(2)平均値の区間推定(大標本の場合)

工場の平均作業時間，製品の平均寿命，食品の有害物質含有量，等々の推定

$\bar{x}$ の標本分布は $N\left(\mu,\ \frac{\sigma^2}{n}\right)$ であり，$z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ は $N(0,1)$ にしたがう。

$$\therefore \quad Pr\left(-1.64 \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le 1.64\right) = 0.90 \qquad (7.5)$$

$$Pr\left(\bar{x}-1.64\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+1.64\frac{\sigma}{\sqrt{n}}\right) = 0.90 \qquad (7.6)$$

同様にして

$$Pr\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95 \qquad (7.7)$$

$$Pr\left(\bar{x}-2.58\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x}+2.58\frac{\sigma}{\sqrt{n}}\right) = 0.99 \qquad (7.8)$$

・母集団の標準偏差 σ が未知の場合は，標本標準偏差 $\hat{\sigma}$ を用いる。

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad (7.9)$$

[例題 2]

在庫中のある種の銅線 30 巻の各々の破断強度(単位 : ポンド)を調べたところ，次のような結果が得られた。

563 548 572 583 570 574 581 568 572 578
568 580 574 581 563 549 567 575 579 564
576 583 576 587 573 566 573 591 572 567

この結果から，この種の銅線の平均破断強度を区間推定せよ。

(3)平均値の区間推定(正規母集団から選ばれた小標本の場合)

$\bar{x}$ の標準化変量 $t = \frac{\bar{x}-\mu}{\frac{\hat{\sigma}}{\sqrt{n}}}$ は自由度$(n-1)$の t 分布にしたがうことになる。

t 値が$[-t_{0.05}(n-1),\ t_{0.05}(n-1)]$ の区間内に含まれる確率は 90%である。

$$\therefore\ Pr\left(-t_{0.05}(n-1) \le \frac{\bar{x}-\mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \le t_{0.05}(n-1)\right) = 0.90 \quad (7.10)$$

$$Pr\left(\bar{x} - t_{0.05}(n-1)\frac{\hat{\sigma}}{\sqrt{n}} \le \mu \le \bar{x} + t_{0.05}(n-1)\frac{\hat{\sigma}}{\sqrt{n}}\right) = 0.90 \quad (7.11)$$

同様にして

$$Pr\left(\bar{x} - t_{0.025}(n-1)\frac{\hat{\sigma}}{\sqrt{n}} \le \mu \le \bar{x} + t_{0.025}(n-1)\frac{\hat{\sigma}}{\sqrt{n}}\right) = 0.95 \quad (7.12)$$

$$Pr\left(\bar{x} - t_{0.005}(n-1)\frac{\hat{\sigma}}{\sqrt{n}} \le \mu \le \bar{x} + t_{0.005}(n-1)\frac{\hat{\sigma}}{\sqrt{n}}\right) = 0.99 \quad (7.13)$$

[例題 3]

ある電球会社で製造した 10 個の電球の寿命時間を測定したところ，「2529, 2520, 2516, 2772, 2593, 2592, 2565, 2645, 2561, 2639」のデータを得た。この結果から，この電球会社製造の電球の平均寿命を区間推定せよ。

7.2 検定(test)

(1)比率の検定

比率の検定は，母集団比率(p)がある特定の値(p_0)に等しいといえるかどうかを調べることである。例えば，放送番組の視聴率や特定の政党に対する有権者の支持率などに関する主張が妥当であるかどうかを調べることである。そのためには，まず，無作為に抽出した標本から標本比率$(\hat{p} = x/n)$を求める。ここで，nは標本の数，xは標本の中の関心事象(例えば，視聴者や支持者)の数である。次に，この統計量を検定仮説に採用される母集団比率と比較し，検定仮設を採択するかどうかを決める。この検定では，大標本の場合，比率の標本分布は正規分布に従うことがわかっている。従って，次の「標本比率の標準化変量(z)」は標準正規分布$N(0,1)$に従うので，この値と有意水準によって決まる境界値との関係に基づいて検定仮説の採用可否を決めればいいことになる。

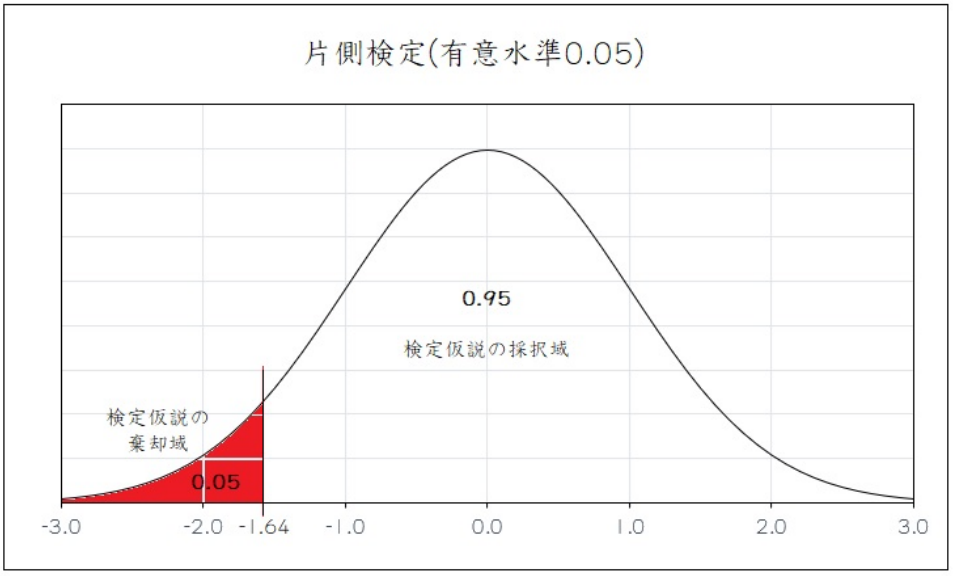
$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \qquad (7.14)$$

例えば，ある選挙区で特定の候補者(A 氏)が自身に対する支持率が少なくとも 75%はあると主張しているとしよう。そこで 1000 人の有権者に対して調査したところ，735 人がＡ氏を支持すると答えたとする。この例における検定仮説と対立仮説(検定仮説が棄却されたときに採用する仮説)は次のように設けられる。

検定仮説　H_0 ： $p = 0.75$

対立仮説　H_1 ： $p < 0.75$

検定仮説が棄却されると対立仮説は「$p < 0.75$」とならざるを得ないので，この例の場合は「片側検定」となる。



有意水準(α)を 0.05(5%)とする場合，検定仮説の棄却域と採択域は上の図のようになり， $z_0 < -1.64$ の場合は検定仮説(H_0)を棄却し，それ以外の場合は検定仮説(H_0)を採択する。標本および仮説からは次のような統計量が得られる。

$$n = 1000,\ \hat{p} = \frac{735}{1000} = 0.735,\ p_0 = 0.75,\ z_0 = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.735-0.750}{\sqrt{\frac{0.75\times0.25}{1000}}} = -1.095$$

$z_0 = -1.095 > -1.64$ であり採択域に入る値であるので，検定仮説(H_0)は採択される。標本比率(0.735)からは A 氏の主張(少なくとも 75%の支持)が妥当でないように思われるが，検定の結果「A 氏の主張は認められる」ことになる。

(2)平均値の検定：正規分布による場合

平均値の検定は，母集団の平均値(μ)がある特定の値(μ_0)に等しいといえるかどうかを調べることである。この場合，母集団がどのような分布であっても標本の数が多ければ正規分布による検定を行うことができる。

まず，母集団の分布が正規分布であり，その分散(σ)が分かっている場合について考察することにしよう。このとき，n個の標本をとり，その平均値を$\bar{x}$とすれば，標本平均値$(\bar{x})$の標準化変量(z)は標準正規分布$N(0,1)$に従うので，この標準化変量の値と有意水準によって決まる境界値との関係に基づいて検定仮説の採用可否を決めればいいことになる。

$$z_0 = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \qquad (7.15)$$

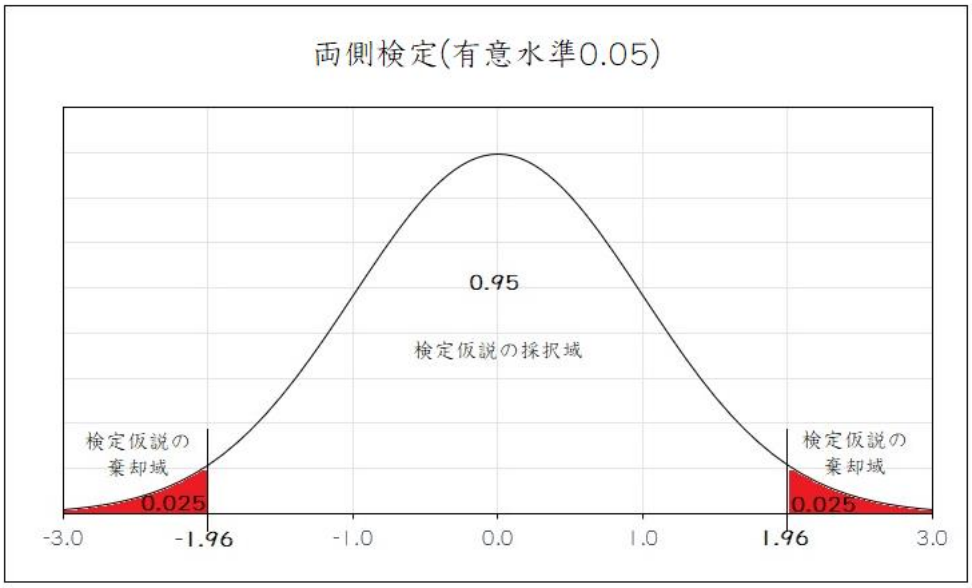
例えば，ある工場で規格直径 10 センチのベアリングを標準偏差 0.3 センチの管理水準で製造しているとしよう。ある日の製造品の中から 10 本の標本をとって直径を測定したところ，平均値が 9.8 センチであった。品質管理上問題なしといえるだろうか。

この例における検定仮説と対立仮説(検定仮説が棄却されたときに採用する仮説)は次のように設けられる。

検定仮説　H_0 ： $\mu = 10$

対立仮説　H_1 ： $\mu \neq 10$

ベアリングは規格より太くても細くてもすべて不良品となるので，検定仮説が棄却された場合に採択される対立仮説は「$\mu \neq 10$」となる。従って，この例の場合は「両側検定」となる。



有意水準(α)を 0.05(5%)とする場合，検定仮説の棄却域と採択域は上の図のようになり， $z_0 < -1.96 \ or \ z_0 > 1.96$ の場合は検定仮説(H_0)を棄却し，それ以外の場合は検定仮説(H_0)を採択する。標本および仮説からは次のような統計量が得られる。

$$n = 10,\ \bar{x} = 9.8,\ \sigma = 0.3,\ \mu_0 = 10,\ z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{9.8 - 10}{0.3/\sqrt{10}} = -2.108$$

$z_0 = -2.108 < -1.96$ であり棄却域に入る値であるので，検定仮説(H_0)は棄却される。従って，「この工場のベアリングは規格通りに造られていない」ことになる。

(3)平均値の検定：t分布による場合

標本の数が少なく$(n \leq 20)$，正規分布に従う母集団の分散(σ^2)が未知の場合，その推定値として標本分散$(\hat{\sigma}^2)$を用いるとき，次の統計量

$$t_0 = \frac{\bar{x} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}} \qquad (7.16)$$

が自由度$\mathrm{n} - 1$のt分布に従うという性質を用いて検定を行うことができる。

前項におけるように正規分布ではなく，t分布を用いることによって生じる相違点は，正規分布の場合に用いられる 1.64(片側検定，α=0.05)，1.96(両側検定，α=0.025)という数値の代わりに，それぞれ t分布の $t_{0.05}(d.f.)$，$t_{0.025}(d.f.)$の値を用いることである。

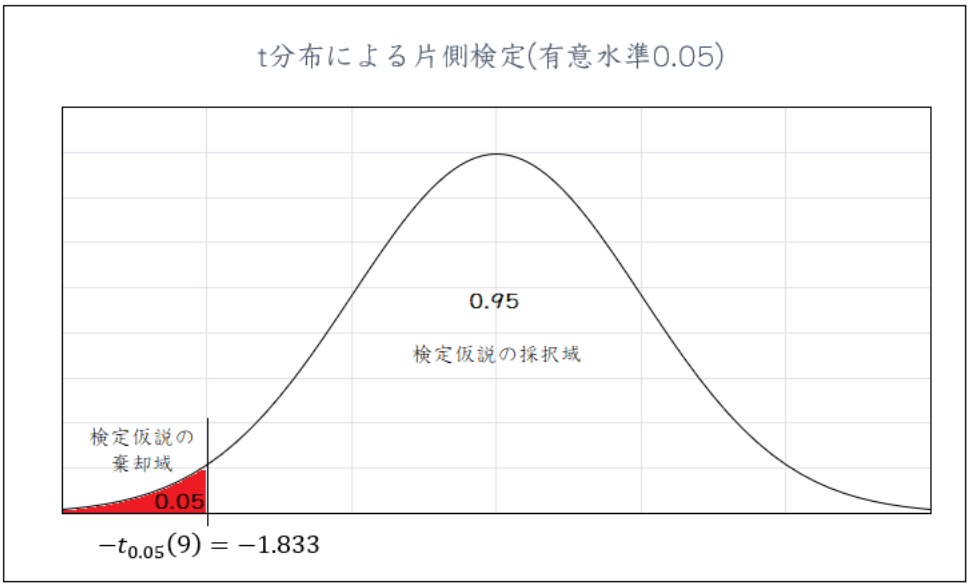
例えば，ある自動車メーカーが自社生産の小型乗用車の燃費について，1 リットル当たり 15km 走行できるとしている。そこで 10 台の実験車両について決められた状態の下で走行テストを行ってみたところ，平均 14.7km，標準偏差 0.5km という結果が得られた。この結果から，この自動車メーカーの主張を認めてよいであろうか。

この例における検定仮説と対立仮説(検定仮説が棄却されたときに採用する仮説)は次のように設けられる。

検定仮説　H_0 ： $\mu = 15$

対立仮説　H_1 ： $\mu < 15$

検定仮説が棄却された場合に採択される対立仮説は「$\mu < 15$」となるので，この例の場合は「片側検定」となる。



有意水準(α)を 0.05(5%)とする場合，検定仮説の棄却域と採択域は上の図のようになり，$t_0 < -t_{0.05}(9) = -1.833$ の場合は検定仮説(H_0)を棄却し，それ以外の場合は検定仮説(H_0)を採択する。標本および仮説からは次のような統計量が得られる。

$$n = 10,\ \bar{x} = 14.7,\ \hat{\sigma} = 0.5,\ \mu_0 = 15,\ t_0 = \frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}} = \frac{14.7-15}{0.5/\sqrt{10}} = -1.897$$

$t_0 = -1.897 < -t_{0.05}(9) = -1.833$ であり棄却域に入る値であるので，検定仮説(H_0)は棄却される。従って，「この小型乗用車の１リットル当たりの燃費は15kmに満たない」という対立仮説が採択されることになる。

[練習問題 1] 母集団比率の区間推定

喫煙者 400 人をランダムに選んで，ある銘柄のタバコを一番好む人の数を調べたところ，95 人であった。喫煙者全体の中でその銘柄を一番好む人の割合 p を区間推定せよ。

[練習問題 2] 母集団平均の区間推定(大標本の場合)

分散$\sigma^2 = 25$の母集団から $n = 200$の標本を取り出したところ，その平均が 20 であった。母集団平均を区間推定せよ。

[練習問題 3] 母集団平均の区間推定(小標本の場合)

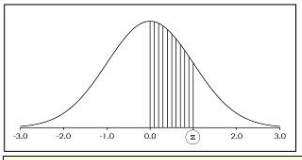
ある機械装置を組み立てるテストを 6 回やったところ、それぞれ 12,13,17,13,15,14 分かかった。この装置を組み立てるのに必要な平均時間を区間推定よ。

[練習問題 4]

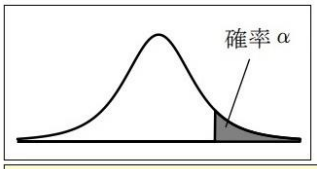
ある電球製造会社は自社製造電球の平均寿命が 720 時間であり，母集団の標準偏差は 20 時間であると公表している。そこで，電球の平均寿命に対するこの会社の公表値の真偽を確かめるために 64 個を電球をランダムに抽出してその平均値を測ってみたところ 715 時間であった。電球の平均寿命に対するこの会社の公表値は信じるに値するであろうか。5%の有意水準で検定せよ。

[練習問題 5]

ある清涼飲料の自動販売機の場合，1 ビン当たり中身が 180cc と表示されている。この自動販売機から無作為に9本を取り出して検査したところ，ビンの中身の平均は 178cc で，標準偏差は 3cc であった。この自動販売機のビンの中身は表示を満たしていると言えるかを 10%の有意水準で検定せよ。



標準正規分布表										
Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
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
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999



片側t分布表					
自由度	α=0.1	α=0.05	α=0.025	α=0.01	α=0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
80	1.292	1.664	1.990	2.374	2.639
120	1.289	1.658	1.980	2.358	2.617
180	1.286	1.653	1.973	2.347	2.603
240	1.285	1.651	1.970	2.342	2.596
∞	1.258	1.645	1.960	2.326	2.576

張本　浩(ハリモト　ヒロシ)
1953 年生まれ。
一橋大学大学院 商学研究科修了
東京国際大学 商学部 教授(管理工学専攻)

統　計　学　入　門

発行日：　2024 年 3 月 1 日
著　者：　張本　浩(harimoto@tiu.ac.jp)
(Tel. 050-3536-0169 内線 5591)