2　代表値

2.1　平均値・中央値・最頻値

分布の中心の位置を表す数値のことを**代表値**といい，平均値，中央値，最頻値等がある。次のデータは男性 25 人の体重である。

59	64	58	68	51	63	57	66	55	54	64	56	74	64	68	
61	56	59	64	59	73	65	57	61	69						

平均値（mean）　$x_1, x_2, \cdots, x_n$ の和を n で割った値を平均値といい，$\bar{x}$ と表す（x バーと読む）。

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$
$$\bar{x} = \frac{59 + 64 + \cdots + 69}{25} = \frac{1545}{25} = 61.8$$

中央値（median）　$x_1, x_2, \cdots, x_n$ は小さいほうから順に並べられているとする。

51	54	55	56	56	57	57	58	59	59	59	61	61	63	64	
64	64	64	65	66	68	68	69	73	74						

中央の順位にあたる値のことを中央値または中位数といい，$\tilde{x}$ と表す。（x チルダと読む）。

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & n \text{ が奇数のとき} \\ \left(x_{\frac{n}{2}} + x_{\frac{n+2}{2}}\right) \div 2 & n \text{ が偶数のとき} \end{cases}$$
$$\tilde{x} = x_{13} = 61$$

最頻値（mode）　最も頻度の高い値のことを最頻値または並数という。最頻値は複数存在することもある。上のデータの最頻値は 64 である。

平均値・中央値・最頻値の大小　周囲より頻度が高まっている部分を峰といい，峰を一つだけもつ分布を単峰（unimodal）という。峰が左にあり，裾が右に厚い分布を「右に歪んだ分布」，その反対を「左に歪んだ分布」という。単峰で右に歪んだ分布では Mode < Median < Mean となる傾向があり，単峰で左に歪んだ分布では Mean < Median < Mode となる傾向がある。

外れ値（outlier）　他の多くの値よりも極めて大きいか極めて小さい値のことを，外れ値という。データに外れ値が含まれていると，平均値はそれに大きく影響されることがある。中央値と最頻値は，外れ値の影響をほとんど受けない。[*i]　*i

[*i] 外れ値の影響を受けにくい性質のことを頑強または頑健（robust）という。

2.2　算術平均・幾何平均・調和平均

算術平均（arithmetic mean）　$x_1, x_2, \cdots, x_n$ の和を n で割った値のことを算術平均あるいは簡単に平均という。

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

$\sum_{i=1}^{n}$ は，$i = 1$ から $i = n$ まで，i の値を 1 ずつ増やしながら，後続の式を加えるための記号である。

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$
$$a\sum_{i=1}^{n} x_i = a\left(x_1 + x_2 + \cdots + x_n\right)$$
$$\sum_{i=1}^{n}(x_i + y_i) = (x_1 + y_1) + \cdots + (x_n + y_n)$$

幾何平均（geometric mean）　$x_1, x_2, \cdots, x_n$ はすべて正数とする。これらの積の（正の）n 乗根のことを幾何平均という。$1/n$ 乗は n 乗根を表している。

$$G = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_n}, \qquad G = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

$\prod_{i=1}^{n}$ は，$i = 1$ から $i = n$ まで，i の値を 1 ずつ増やしながら，後続の式を掛けるための記号である。

$$\prod_{i=1}^{n} x_i = x_1 \times x_2 \times \cdots \times x_n$$

次の表はある銘柄について，前日の価格と比較した増減率を示したものである。

日付：	1	2	3
増減：	$+80\%$	-40%	$+60\%$

比率の平均値は幾何平均 G によって求める。

$$G = \sqrt[3]{(1+0.8)\times(1-0.4)\times(1+0.6)} = \sqrt[3]{1.728} = \sqrt[3]{(1.2)^3} = 1.2$$

1 日当たりの平均増加率は $+20\%$ となる。

調和平均（harmonic mean）　$x_1, x_2, \cdots, x_n$ はすべて正数とする。これらの逆数の算術平均の逆数のことを調和平均という。-1 乗は逆数を表している。

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}, \qquad H = \left(\frac{1}{n}\sum_{i=1}^{n} {x_i}^{-1}\right)^{-1}$$

片道 180 km の距離を往路は 60 km/h，復路は 90 km/h で走るとき，往復の平均速度 H を求める。

	往	復	往復
距離：	180	180	360
時間：	3	2	5
速度：	60	90	H

360 km を 5 時間で走るから往復の平均速度は 72 km/h である。

$$H = \frac{180+180}{3+2} = \frac{180+180}{\frac{180}{60}+\frac{180}{90}} = \frac{180\,(1+1)}{180\left(\frac{1}{60}+\frac{1}{90}\right)} = \frac{2}{\frac{1}{60}+\frac{1}{90}}$$

60，90 の調和平均は 72 になる。

算術平均・幾何平均・調和平均の大小　算術平均 A，幾何平均 G，調和平均 H について，次の不等式が常に成立することが知られている。[*ii] [*ii]

$$A \geqq G \geqq H$$

2.3　その他の代表値

刈り込み平均（trimmed mean）　最小側と最大側の一定数を除外した残りのデータの平均値のこと。除外する個数や割合に決まりはない[*iii]。例えば，最小側の r 個，最大側の r 個を除外した刈り込み平均は [*iii]

$$\text{Trimmed mean} = \frac{x_{r+1} + x_{r+2} + \cdots + x_{n-r-1} + x_{n-r}}{n-2r}$$

となる。ただし $x_1, x_2, \cdots, x_n$ は小さいほうから順に並べられているとする。刈り込み平均は，採点競技等でよく利用される。

ミッドレンジ（midrange）　最小値 x_1 と最大値 x_n の平均値のこと。

$$\text{Midrange} = \frac{x_1 + x_n}{2}$$

[*ii] 不等式 $A \geq G \geq H$ は凸関数の性質等から導かれる。
[*iii] 除外するデータを多くすると，刈り込み平均は中央値に近づく。

2.4 度数分布表から代表値を求める

次の表は男性 25 人の体重から作った分布表である。

階級	階級値	度数	相対度数	累積度数	累積相対度数
$47.5 \sim 52.5$	50	1	0.04	1	0.04
$52.5 \sim 57.5$	55	6	0.24	7	0.28
$57.5 \sim 62.5$	60	6	0.24	13	0.52
$62.5 \sim 67.5$	65	7	0.28	20	0.80
$67.5 \sim 72.5$	70	3	0.12	23	0.92
$72.5 \sim 77.5$	75	2	0.08	25	1.00

平均値　階級値を x_j，度数を f_j，度数の総和を n，相対度数を $p_j = f_j/n$ とする。

階級	階級値	度数	相対度数
$a_0 \sim a_1$	x_1	f_1	p_1
$a_1 \sim a_2$	x_2	f_2	p_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$a_{k-1} \sim a_k$	x_k	f_k	p_k
計	–	n	1

階級 $a_{j-1} \sim a_j$ に含まれるすべての値を階級値 x_j に置き換えることによって，平均値を近似することができる[*iv]。次の (1) を度数 f_j を重みとする**加重平均**または重み付き平均（weighted mean）という。

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}{f_1 + f_2 + \cdots + f_k} = \frac{1}{n}\sum_{j=1}^{k} x_j f_j \tag{1}$$

$$\bar{x} = \frac{50 \cdot 1 + 55 \cdot 6 + \cdots + 75 \cdot 2}{1 + 6 + \cdots + 2} = \frac{1555}{25} = 62.2$$

次の (2) は相対度数 p_j を重みとする加重平均である。(1) と (2) の値は等しい。

$$\bar{x} = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k = \sum_{j=1}^{k} x_j p_j \tag{2}$$

$$\bar{x} = 50 \cdot 0.04 + 55 \cdot 0.24 + \cdots + 75 \cdot 0.08 = 62.2$$

最頻値　度数が最も大きい（相対度数が最も大きい）階級の階級値を最頻値とする。最頻値は複数存在することもある。上のデータの最頻値は 65 である。

中央値　累積度数がはじめて $n/2$ を超える（累積相対度数がはじめて 0.5 を超える）階級の階級値を中央値とする[*v]。上のデータの中央値は 60 である。

[*iv] 階級値への置き換えによる誤差は,「階級の幅」の半分までに抑えられる。

[*v] 階級の境界で累積度数が $n/2$ になる（累積相対度数が 0.5 になる）場合は，階級の境界値を中央値とする。

2.5　仮平均法

例　変数 x が，2011，2008，2006，2001，1994 のとき，$y_i = x_i - 2000$ とおくと，変数 y は，11，8，6，1，-6 だから，$\bar{y} = 4$ となる。もとの変数 x の平均値は，$\bar{x} = 2000 + 4 = 2004$ である。

例　変数 x が，13.07，13.23，13.31，13.66，13.78 のとき，$y_i = 100(x_i - 13)$ とおくと，変数 y は，7，23，31，66，78 だから，$\bar{y} = 41$ となる。もとの変数 x の平均値は，$\bar{x} = 13 + 41 \div 100 = 13.41$ である。

変数の変換　もとの変数 x から新しい変数 y を，$y_i = ax_i + b$ のように定めると，その平均値について，$\bar{y} = a\bar{x} + b$ が成り立つ。

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

特に $y_i = (x_i - c)/d$ のとき，$x_i = c + dy_i$ と変形できるから，次のことが成り立つ。

$$y_i = \frac{x_i - c}{d},\ x_i = c + dy_i \implies \bar{x} = c + d\bar{y} \tag{3}$$

仮平均（assumed mean）　もとの変数 x から新しい変数 y を $y_i = (x_i - c)/d$ と定めると，$\bar{x} = c + d\bar{y}$ が成り立つ。c の値のことを仮平均という。

1 番目の例では仮平均等を $c = 2000$，$d = 1$ として，2 番目の例では仮平均等を $c = 13$，$d = 1/100$ として，もとの変数の平均値を求めている。

変数の変換により平均値を求める　度数分布表から平均値を求める場合，c に階級値のいずれか，d に階級の幅を代入すると y_j が簡単な値になる。

階級	階級値 x	y	度数	相対度数
$47.5 \sim 52.5$	50	-3	1	0.04
$52.5 \sim 57.5$	55	-2	6	0.24
$57.5 \sim 62.5$	60	-1	6	0.24
$62.5 \sim 67.5$	65	0	7	0.28
$67.5 \sim 72.5$	70	1	3	0.12
$72.5 \sim 77.5$	75	2	2	0.08

$c = 65$（最頻値），$d = 5$（階級の幅），$y_j = (x_j - 65)/5$ とおくと，

$$\bar{y} = \frac{(-3)\cdot 1 + (-2)\cdot 6 + \cdots + 2\cdot 2}{25} = \frac{-14}{25} = -0.56$$

$$\bar{x} = 65 + 5\cdot\bar{y} = 60 + 5\cdot\frac{-14}{25} = 62.2$$

参考文献

統計学入門（基礎統計学）
東京大学教養学部統計学教室（編）東京大学出版会 978-4-13-042065-5

統計学
久保川 達也 （著） 東京大学出版会 978-4-13-062921-8

はじめての統計学
道家 暎幸（共著） コロナ社 978-4-339-06113-0

確率統計 新版 （新版数学シリーズ）
岡本 和夫 （ほか著） 実教出版 978-4-407-32171-5

統計学序論 改訂版
山本 義郎 （著） 東海大学出版部 978-4-486-02133-9

確率統計（高専テキストシリーズ）
上野 健爾（監修） 森北出版 978-4-627-05561-2

基本統計学 第４版
宮川 公男 （著） 有斐閣 978-4-641-16455-0

新統計入門
小寺 平治 （著） 裳華房 978-4-7853-1099-8

Schaum's Outline of Introduction to Probability and Statistics
Seymour Lipschutz （著） McGraw-Hill Education 978-0-07-176249-6

A Dictionary of Statistics
Graham Upton （著） Oxford Univ Pr 978-0-19-967918-8

www5e.biglobe.ne.jp/~emm386/statistics/