記述統計 [データ解析]

石綿 元

第 1 回講義

目次

Copyright GEN ISHIWATA 石綿 元

1　データの視覚化

1.1　統計が扱うデータ

1.1.1　データの種類

■質的データ　分類の属性や状態を区分したデータ

例）性別、地域、天候、etc

■量的データ　単位を用いて数値で表せるテータ

例）長さ、速さ、重さ、etc

1.1.2　データの次元

■一次元データ　一つの観測対象に対して一つの観測値が与えられているデータ

例）学生と身長

■二次元データ　一つの観測対象に対して二つの観測値が与えられているデータ

例）学生と身長と体重

■多次元データ　一つの観測対象に対して多数の観測値が与えられているデータ

例）アンケート

1.1.3　クロスセクションデータ

同一時点での異なった複数の項目について観測値を並べたデータ

例）都県別の人口と世帯数 (2016 年 8 月時点での各都県公式ホームページ公表データ)

都県	群馬	栃木	茨城	埼玉	千葉	東京	神奈川
人口 [人]	1,983,581	1,974,720	2,917,857	7,261,271	6,224,027	13,617,445	9,146,941
世帯数 [世帯]	755,756	762,535	1,123,802	2,968,978	2,607,079	6,788,652	4,017,683

1.1.4　連続型変数と離散型変数

■連続型変数　観測値が実数として得られる場合の変数

例）学生と身長、体重など

■離散型変数　観測値が整数として得られる場合の変数

例）世帯の子供の数、不良品の個数など

1.2　データを記述する

集めてきた量的データは数字の羅列であり、そこから意味のある情報を取り出す。データから、有効な情報を読み取るために

- 表を作る。

Copyright GEN ISHIWATA 石綿 元

- グラフを作る。
- 度数分布表を作る。
- ヒストグラムを作る。
- 特徴付ける値を計算により求める。

などを行う。これを**データを記述する**という。

1.2.1 度数分布表に分類する

■度数分布表とは データの散らばりを表すための表で、特定の範囲の中にいくつのデータが含まれるか数を表にしたものである。

■度数分布表の作り方

1. レンジを求める。
   - レンジとは、値の存在する範囲のことである。

$$\text{レンジ} = \text{最大値} - \text{最小値} = \text{範囲}$$

2. 階級幅を決める。
   - 階級とは、データをまとめる幅のことである。
   - 階級を 10 前後に決めるならば、レンジを 10 で割った値とする。
3. 階級を決定する。
   - 測定単位を考慮して階級の下側限界値を決める。
4. 度数を数える。
   - 各階級に含まれるデータの個数を数える。

データ例）

158.0	166.0	168.0	170.0	170.5	173.0	176.0	179.0
160.0	166.0	168.1	170.0	170.6	173.0	176.2	179.5
161.0	166.0	168.2	170.0	171.0	173.1	176.3	179.7
162.0	166.2	168.2	170.0	171.0	173.5	176.8	180.0
162.0	166.2	168.4	170.0	171.0	173.8	177.0	180.0
162.1	166.6	168.5	170.0	171.0	174.0	177.0	180.0
162.6	167.0	168.8	170.0	171.2	174.1	177.4	180.1
163.0	167.0	168.9	170.0	171.3	174.7	177.8	180.6
163.0	167.0	169.0	170.0	171.4	174.8	178.0	180.6
163.0	167.0	169.0	170.0	171.5	174.8	178.0	180.6
163.1	167.0	169.0	170.0	172.0	175.0	178.0	180.8
163.9	167.0	169.1	170.0	172.0	175.0	178.0	181.0
164.0	167.0	169.4	170.1	172.0	175.0	178.0	181.0
164.2	167.3	169.6	170.1	172.0	175.0	178.0	181.5
164.3	167.5	169.7	170.1	172.2	175.0	178.0	182.0
164.9	167.5	169.7	170.3	172.2	175.0	178.0	182.4
165.0	167.7	169.8	170.4	172.4	175.5	178.0	182.6
165.0	168.0	169.8	170.4	172.5	175.7	178.4	184.0
165.0	168.0	169.8	170.4	172.6	175.7	178.8	185.0
165.0	168.0	170.0	170.4	173.0	176.0	179.0	187.2

図１　大学生（男子）１６０人の身長

このデータを度数分布表にまとめる。

Copyright GEN ISHIWATA 石綿 元

階級を下側限界値 150[cm] から階級幅を 5[cm] づつにとると、次の度数分布表が得られる。

階級	度数
150.0から154.9	0
155.0から159.9	1
160.0から164.9	15
165.0から169.9	43
170.0から174.9	51
175.0から179.9	33
180.0から184.9	15
185.0から189.9	2
190.0から194.9	0

図 2　度数分布表

1.3　ヒストグラムで見る

1.3.1　度数分布表からヒストグラムを作る

度数分布表に分類したデータをヒストグラムでみる。

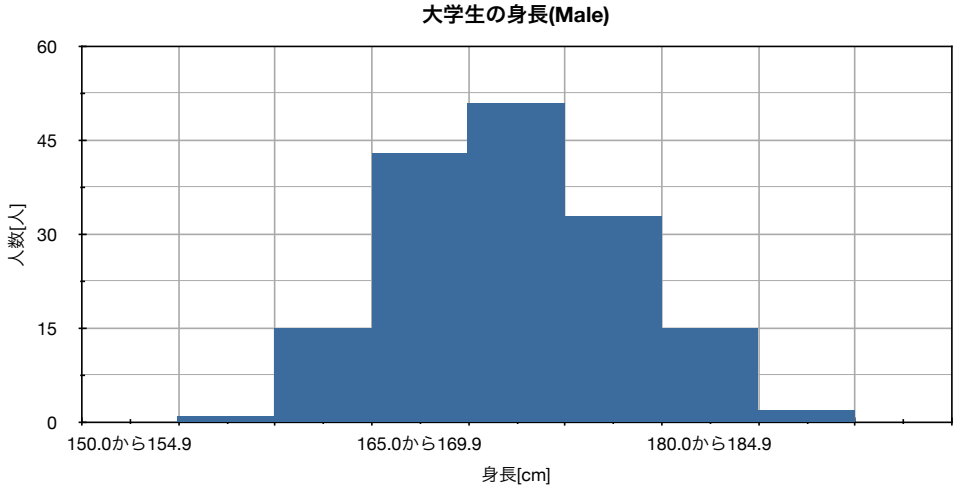

図 3　ヒストグラム

1.3.2　ヒストグラムの特徴をとらえる

1. 単峰型（一山型）であるか。
   - 多くの場合、単峰型（一山形）になる。単峰でないならば、その原因を探る。
2. 左右対称であるか。
   - たいていの場合、左右対称となることが多い。
3. はずれ値はあるか。
   - はずれ値があるならば、その値がどのような意味を持つのか、持たないのか考察する。

Copyright GEN ISHIWATA 石綿 元

1.3.3　度数分布表の拡張

階級	階級値	度数	相対度数	累積度数	相対累積度数
150.0から154.9	152.45	0	0	0	0
155.0から159.9	157.45	1	0.00625	1	0.00625
160.0から164.9	162.45	15	0.09375	16	0.1
165.0から169.9	167.45	43	0.26875	59	0.36875
170.0から174.9	172.45	51	0.31875	110	0.6875
175.0から179.9	177.45	33	0.20625	143	0.89375
180.0から184.9	182.45	15	0.09375	158	0.9875
185.0から189.9	187.45	2	0.0125	160	1
190.0から194.9	192.45	0	0	160	1
		160	1		

図 4　相対度数と相対累積度数

■度数と相対度数　幅を持たせて決定した階級について、その階級を代表する値を階級値とよび、その階級幅の中央の値を当てる。相対度数とは、各階級に含まれる度数を総数で割って各階級における度数の割合に変えた値である。度数を割合に変えることにより各階級に属するデータの出現確率として読み取ることができる。

■累積度数と相対累積度数　累積度数とは、その階級以下の階級までに含まれる度数の和で、各階級の度数を累積して表したものである。相対累積度数とは、各階級に含まれる累積度数を総数で割って各階級における累積度数の割合に変えた値である。

1.3.4　相対累積度数折れ線グラフで読む

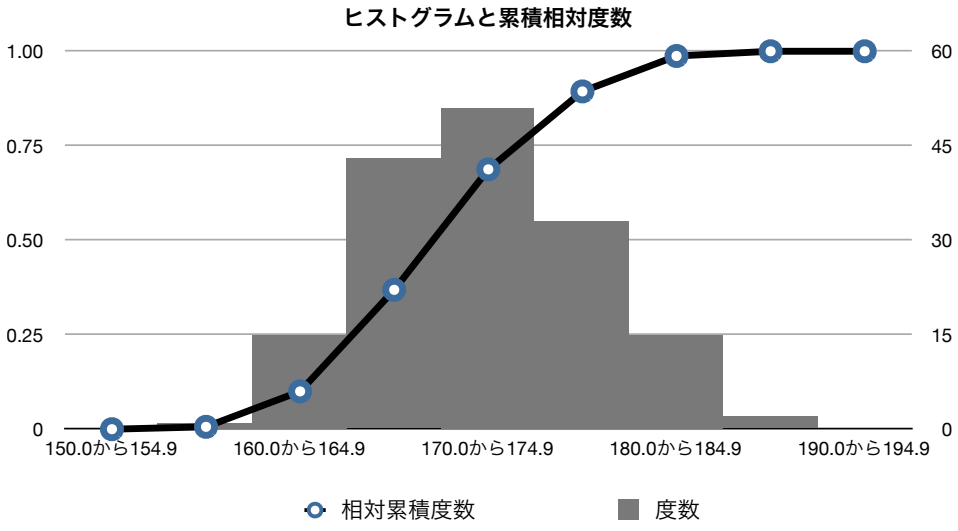

図 5　ヒストグラムと相対累積度数折れ線グラフ

相対累積度数を求めることで、全体の何％までの値（階級）を求めたいときや、ある階級までは、全体の何％であるかを求めたいときなどに効果的である。

Copyright GEN ISHIWATA 石綿 元

1.3.5 層別

例えば、先の例のように大学生の背の高さについてヒストグラムを作ることにしよう。先の例では、男子のみのデータを用いてヒストグラムを作成したが、男女同数の混合データについてヒストグラムを作ると、単峰型にはならない。山は２つ現れてくる。なぜならば、男子と女子では、平均身長が異なるからである。このようなときは、男子のヒストグラムと女子のヒストグラムを別々に作ることで一山のヒストグラムにできる。このように複峰のヒストグラムについて峰ごとにそうなる理由を考察し、ヒストグラムを分けることを層別と呼ぶ。

1.4 その他の代表的な図表

1.4.1 円グラフ

円グラフは、円の中で、各項目の構成比が面積で示され、一目で調査対象の大小関係がわかるように配置したものである。そのために、単独項目のデータは大きい順に時計回りに配置し、その他の項目は常に最後に配置する。

データ例）全世界の出荷ノートパソコンのメーカーシェアのデータとグラフ[*1]

メーカー	HP	Lenovo	Dell	Apple	ASUS	acer	TOSHIBA	Samsung	Others
マーケットシェア	20.5	19.9	13.7	10.34	10.31	8.9	4.2	1.7	10.45

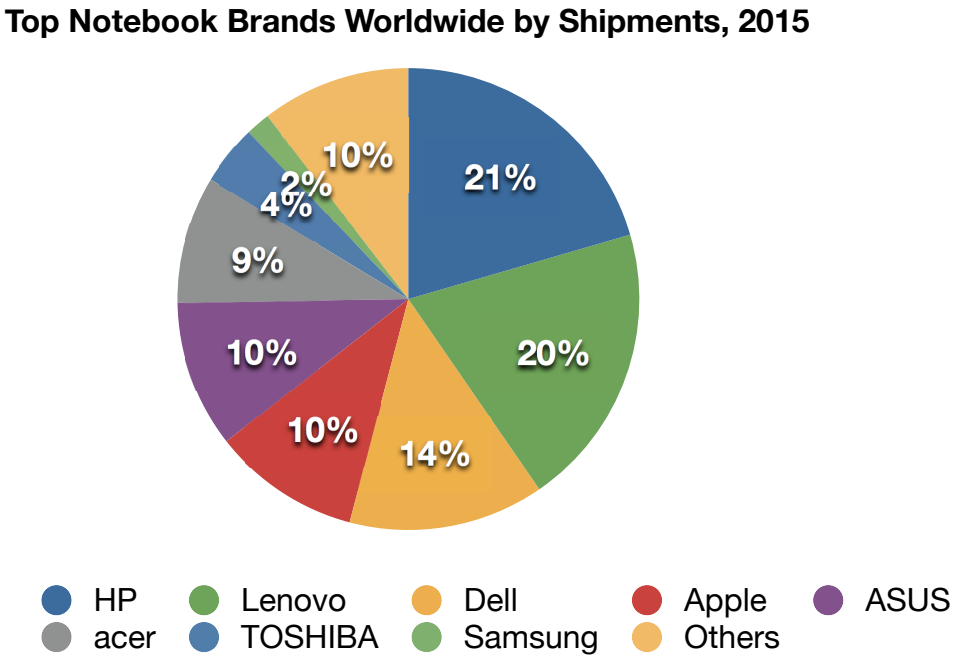

図 6 円グラフ

[*1] データ出典：http://press.trendforce.com/press/20160216-2327.html

Copyright GEN ISHIWATA 石綿 元

1.4.2　散布図

二つのデータに関係があるかどうかを調べる場合に使われるのが散布図である。x 軸と y 軸で別の量を変数に取り、双方の値の交わる点にプロットしていく。

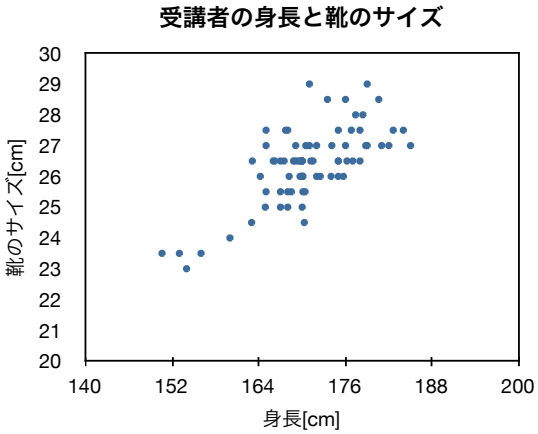

図 7　散布図

1.4.3　各グラフの特徴と使い分け

表現したいデータが得られたならば、目的ごとにグラフは使い分けなければならない。グラフによって表現できる内容が異なり、データの特徴にあったグラフで表現することが必要である。

■表現目的とそれが得意な主なグラフの対応

データの大小を比較する　棒グラフ
データの散らばりを見る　ヒストグラム
データの散らばりを比較する　箱ひげ図
データの構成比を見る　円グラフ
データの構成比を比較する　帯グラフ
２つのデータの関係を見る　散布図

棒グラフは、その名の通りデータを棒で表現する最も単純でよく使われるグラフである。棒の高さにより、値の大小を表現し、グラフ全体の値について比較ができる。

ヒストグラムは、一見すると棒グラフに似ているが、各階級は範囲を持っており、各棒は高さだけでなく幅も意味を持っている。したがって、面積にも意味がある。単一データセットの散らばりを見るのに適している。複数のデータセットについての散らばり具合の比較を行う場合は、箱ひげ図を用いる。箱ひげ図については、「2　データを代表値で読む」で扱う。

円グラフは、円の中の面積によって、各項目の全体に占める割合を視覚的に表現することが得意なグラフである。複数のデータセットについて各項目の構成比を表現したい場合は、円を複数描くことよりも帯グラフが良い。帯グラフは、帯の中に面積によって各項目の全体に占める割合を表現するが、同じ長さの複数の帯を並べることによって、構成比を比較することができるようにしたものである。

Copyright GEN ISHIWATA 石綿 元

2 データを代表値で読む

2.1 記述的尺度 I

2.1.1 位置の尺度

■算術平均 定義：n 個の点でのデータの総和を個数 n で割る。

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

度数分布表のみが与えられている場合、データの総数 n で、r 個の各階級に属する度数を m_i、各階級の階級値を x_i として

$$\bar{x} = \frac{\sum_{i=1}^{r} \mathrm{x}_i m_i}{\sum_{i=1}^{r} m_i} = \frac{\sum_{i=1}^{r} \mathrm{x}_i m_i}{n} = \frac{1}{n}\sum_{i=1}^{r} \mathrm{x}_i m_i = \sum_{i=1}^{r} \mathrm{x}_i \frac{m_i}{n}$$

算術平均とは、そのデータでヒストグラムを作成したときの**重心**を表している。このことは、ヒストグラムそのものが横軸上でどこに位置するのかを決める量である。

■最頻値（モード） 観測値の組の中で、最大の度数をもつ観測値をモードまたは最頻値と定義する。

データ例）3, 3, 4, 4, 4, 5, 5, 6, 6, 7, 8, 9, 9

値（データ）	3	4	5	6	7	8	9
個数（度数）	2	**3**	2	2	1	1	2

この場合、最大の度数は 3 個である。よって、モードは 4 である。

■中央値（メディアン） 観測値を大きさの順に並べたときのちょうど真ん中に来る値をメディアンまたは中央値と定義する。

データ例）3, 3, 4, 4, 4, 5, 5, 6, 6, 7, 8, 9, 9

値（データ）	3	3	4	4	4	5	5	6	6	7	8	9	9
順番（順位）	1	2	3	4	5	6	**7**	8	9	10	11	12	13

この場合、データは 13 個あるので、中央は第 7 番目。よって、メディアンは 5 である。

■百分位数 ヒストグラムの面積全体を 100 として観測値が整数で何パーセントの位置にあるかを表す。つまり、ヒストグラムの面積を百等分する点。

■四分位数 ：ヒストグラムの面積を四等分する次の 3 点をまとめて四分位数という。

第 1 四分位数 データを小さい順に並べて下から $\frac{1}{4}$ の場所の値。
第 2 四分位数 中央値と同じ。データを並べて $\frac{1}{2}$ の場所の値
第 3 四分位数 データを小さい順に並べて下から $\frac{3}{4}$ の場所の値。

四分位範囲

$$\text{四分位範囲} = \text{第 3 四分位数} - \text{第 1 四分位数}$$

Copyright GEN ISHIWATA 石綿 元

2.2 記述的尺度 II

2.2.1 散らばりの尺度

■偏差 偏差とは、そのデータの平均からの各データまでの距離 $x_i - \bar{x}$ をいう。

■分散 平均値、最頻値や中央値などでデータを代表する値を計算できても、データを十分に説明できない場合がある。例えば、ヒストグラムを作成したときに、純粋な単峰型かつ完全な左右対称の場合、平均値も最頻値も中央値もいずれも同じ値を取ることになる。このような性質を持ったデータの場合、複数の異なるデータセットが、いずれのデータセットでも平均値が同じ値ならば、最頻値も中央値も同じ値となる。つまり、平均値、最頻値、中央値などの位置の尺度においては区別がつかない。しかし、データを代表値で表現するという目的からは、違うデータセットなのにそれらの区別がつかないのは不完全である。そこで新たな尺度を導入する。

たとえば、二つの異なるデータセットがあって平均値も最頻値も中央値もいずれも同じ値を取るデータセットであってもレンジが違う場合はあり得る。つまり、平均値も最頻値も中央値も同じ値であってもその値の周辺にどの程度データが集中しているのかが違うということである。そこで、偏差を用いてその平均を考えることにする。一見、良い尺度のように思えるが、もし、平均値が 0 ならば、左右対称なデータの偏差の平均は、平均値の右側の各データまでの距離と平均値の左側の各データまでの距離が等しい訳だから、この値は正負の符号が逆転して同数のため常に 0 である。これではやはりこの二つのデータセットの区別ができない。

そこで、正負の符号の影響を排除するために偏差を二乗することにする。これを偏差平方といい、偏差平方の平均を分散と定義する。

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

度数分布表のみが与えられている場合、r 個の各階級に属する度数を m_i、各階級の階級値を x_i として

$$s^2 = \frac{1}{\sum_{i=1}^{r} m_i}\sum_{i=1}^{r}(\mathrm{x}_i - \bar{x})^2 m_i = \frac{1}{n}\sum_{i=1}^{r}(\mathrm{x}_i - \bar{x})^2 m_i = \sum_{i=1}^{r}(\mathrm{x}_i - \bar{x})^2 \frac{m_i}{n}$$

■標準偏差 偏差平方の平均を分散と定義したが、二乗しているので、データが持つ単位との整合性を考えれば、根号を付けて戻しておく。これを標準偏差と定義する。

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

2.3 箱ひげ図

■五数要約 データのばらつきについて、データの最小値・四分位数・最大値の全部で 5 つの数で表すことを五数要約という。

■箱ひげ図 五数要約の 5 つの値を用いて、最小値と最大値を線で結び、第 1 四分位数と第 3 四分位数を箱で表し、箱の中に中央値を表すグラフを箱ひげ図と呼ぶ。データの範囲とばらつきを読み取りやすくしたものである。

ヒストグラムを用いればデータの散らばりを見ることができたが、複数のグラフでデータの散らばりを比較するには同じグラフ上で並べて表示できる箱ひげ図がヒストグラムよりも優れている。

Copyright GEN ISHIWATA 石綿 元

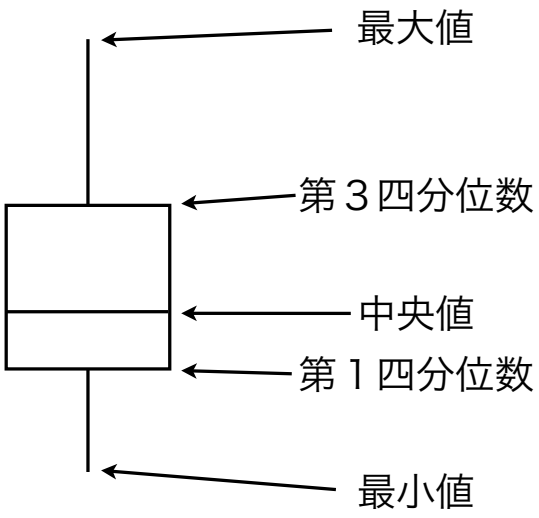

図 8　箱ひげ図

2.4　時系列データ

2.4.1　移動平均

観測値が季節変動などによって短い周期で振動している場合、全体的な大きな変動が背後に隠れていても、とらえることが難しい。そのようなときには、移動平均を用いることで、データ変動が平滑化され、大きな視野でデータをみることができる。

■四季による周期のあるデータの移動平均　日本には四季がある。そのため例えば小売業において衣料品や家電などの売り上げは季節変動が激しい。月ごとに売り上げデータを見て前月と比較してもあまり意味がない。そこで、移動平均を計算してデータを平滑化しておくと、その業態の売り上げが徐々に伸びているか、全体的に衰退しているのかの判断ができる。

$$M_{A_4} = \frac{1}{4}\left(\frac{1}{2}x_{t-2} + x_{t-1} + x_t + x_{t+1} + \frac{1}{2}x_{t+2}\right)$$

■移動平均　明確な長い周期がわからない場合は、データの特性に応じて５や６など少し大きい数の移動平均を見てみることも意味があるかもしれない。次の例は、気象庁が発表したデータによる平均気温の偏差の変動のそのままのデータでの折れ線グラフと５年移動平均および６年移動平均の折れ線グラフである。元データの振動は大きいが、移動平均で平滑化しておくと、右肩上がりなグラフがより強調されて出てくる。つまり、年ごとに見る平均気温の変動は大きいが、移動平均による平滑化で平均気温が徐々上昇していることがよりわかりやすくなる。

$$M_{A_5} = \frac{1}{5}\left(\frac{1}{2}x_{t-2} + x_{t-1} + x_t + x_{t+1} + x_{t+2} + \frac{1}{2}x_{t+3}\right)$$

$$M_{A_6} = \frac{1}{6}\left(\frac{1}{2}x_{t-3} + x_{t-2} + x_{t-1} + x_t + x_{t+1} + x_{t+2} + \frac{1}{2}x_{t+3}\right)$$

Copyright GEN ISHIWATA 石綿 元

年	偏差	年	偏差	年	偏差	年	偏差	年	偏差	年	偏差
1898年	-0.73	1918年	-1.09	1938年	-0.54	1958年	-0.29	1978年	-0.16	1998年	+0.75
1899年	-0.81	1919年	-0.74	1939年	-0.68	1959年	+0.12	1979年	+0.20	1999年	+0.49
1900年	-1.06	1920年	-0.51	1940年	-0.85	1960年	-0.22	1980年	-0.78	2000年	+0.28
1901年	-1.03	1921年	-1.08	1941年	-0.79	1961年	+0.18	1981年	-0.98	2001年	-0.05
1902年	-1.03	1922年	-0.58	1942年	-0.67	1962年	-0.37	1982年	-0.33	2002年	+0.29
1903年	-0.76	1923年	-0.76	1943年	-0.86	1963年	-0.64	1983年	-0.49	2003年	-0.06
1904年	-0.86	1924年	-1.01	1944年	-1.09	1964年	-0.29	1984年	-0.99	2004年	+0.77
1905年	-0.95	1925年	-0.94	1945年	-1.57	1965年	-0.99	1985年	-0.37	2005年	-0.01
1906年	-1.32	1926年	-1.32	1946年	-0.36	1966年	-0.53	1986年	-0.95	2006年	+0.20
1907年	-1.21	1927年	-0.93	1947年	-1.42	1967年	-0.47	1987年	-0.13	2007年	+0.61
1908年	-1.44	1928年	-0.62	1948年	-0.08	1968年	-0.64	1988年	-0.65	2008年	+0.22
1909年	-1.13	1929年	-0.87	1949年	-0.65	1969年	-0.79	1989年	+0.16	2009年	+0.30
1910年	-1.22	1930年	-0.32	1950年	-0.29	1970年	-0.75	1990年	+0.78	2010年	+0.61
1911年	-0.70	1931年	-1.04	1951年	-0.62	1971年	-0.68	1991年	+0.25	2011年	+0.13
1912年	-1.12	1932年	-0.72	1952年	-0.74	1972年	-0.14	1992年	-0.11	2012年	+0.04
1913年	-1.59	1933年	-0.64	1953年	-0.74	1973年	-0.30	1993年	-0.52	2013年	+0.34
1914年	-0.20	1934年	-1.14	1954年	-0.53	1974年	-0.91	1994年	+0.56	2014年	+0.14
1915年	-0.56	1935年	-0.76	1955年	-0.12	1975年	-0.35	1995年	-0.19	2015年	+0.69
1916年	-0.14	1936年	-1.10	1956年	-0.74	1976年	-0.87	1996年	-0.54		
1917年	-1.31	1937年	-0.37	1957年	-0.76	1977年	-0.41	1997年	+0.10		

図 9　日本の平均気温の偏差：気象庁ホームページより

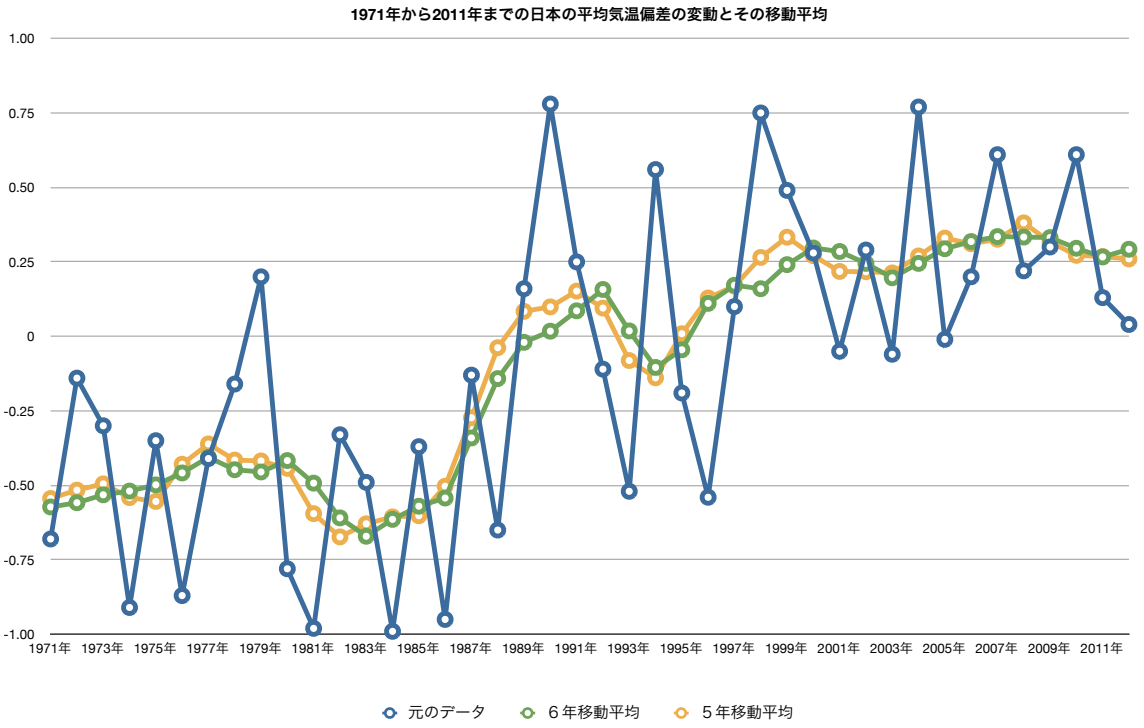

図 10　移動平均

2.4.2　幾何平均

データの変動が前年比など比率で表されている場合は幾何平均を用いる。

例えば、我が社の今年度の売り上げは前年比 1.23 %増である！などという話はよくある話であるが、この前年比 1.23 %という数字は、多くの場合毎年同じではない。では、この会社のここ 5 年間の売り上げ前年比

Copyright
GEN ISHIWATA
石綿 元

のデータを調べてみると、4 年前が 7.7 %増、3 年前が 3.9 %増、2 年前が 4.2 %増、1 年前が 1.8 %増で今年が 1.23 %増であったとしよう。5 年前の売り上げを a 円とすると本年度の売り上げ b 円は、

$$a \times (1.077) \times (1.039) \times (1.042) \times (1.018) \times (1.0123) = b$$

である。いま、もしこの 5 年間一定の比率 x で 5 年前の a 円から本年度の売り上げ b 円と求まるとすると、

$$a \times (x) \times (x) \times (x) \times (x) \times (x) = b$$

である。この x を幾何平均と呼ぶ。つまり、毎年一定の比率にするといくつですか？という意味である。

くだんの例で見ると、

$$a \times (1.077) \times (1.039) \times (1.042) \times (1.018) \times (1.0123) = a \times (x) \times (x) \times (x) \times (x) \times (x)$$

であるから、$x^5 = (1.077) \times (1.039) \times (1.042) \times (1.018) \times (1.0123)$ である。つまり、

$$x = \sqrt[5]{(1.077) \times (1.039) \times (1.042) \times (1.018) \times (1.0123)}$$

すなわち、比率データ n 個の幾何平均は、

$$Gm_n = \sqrt[n]{x_1 \times x_2 \times x_3 \times \cdots \times x_n}$$

Copyright GEN ISHIWATA 石綿 元

索引

Copyright GEN ISHIWATA 石綿 元