統計学入門　第２章 1次元のデータ

張 梁

2007年2月23日

目 次

2　1次元のデータ

正しくしかも効率的にデータを読むには、

記述統計学 (descriptive statistics)　集団としての特徴を記述するために、観測対象となった各個体について観測し、得られたデータを整理・要約する方法である

観測 (observation)　広く調査や実験のことである

データ (data)　観測から、各個体の観測値を得て、それをまとめたものをデータという

一次元のデータ (1-dimensional data)　各個体について得られたデータのうちの1種類だけ

2.1　度数分布とヒストグラム

2.1.1　度数分布表

階級と級数と対応関係を表にしたもの

表 2.1　試験得点の度数分布表(某大学の統計学)

階級		階級値	度数	相対度数	累積度数	累積相対度数
0点以上	10点未満	5	12	0.032	12	0.032
10 〃	20 〃	15	10	0.027	22	0.059
20 〃	30 〃	25	19	0.051	41	0.110
30 〃	40 〃	35	42	0.113	83	0.223
40 〃	50 〃	45	72	0.193	155	0.416
50 〃	60 〃	55	82	0.220	237	0.635
60 〃	70 〃	65	54	0.145	291	0.780
70 〃	80 〃	75	38	0.102	329	0.882
80 〃	90 〃	85	25	0.067	354	0.949
90 〃	100点以下	95	19	0.051	373	1.000
合計			373	1.000		

試験得点の分布において，つねにこのような整った結果が得られるとはかぎらない．

階級 (class)　観測値のとりうる値をいくつかに分ける

階級値　各階級の中では観測値は一様に分布していると仮定する。階級の上限値と下限値の中間値を、階級を代表する値、階級値とする

度数 (frequency)　それぞれの階級で、観測値がいくつあるかを表す

相対度数 (relative frequency)　データ全体の大きさを 1 として、各階級に属する観測値の全体中での割合を示す

累積度数 (cumulative frequency)　度数を下の階級から順に積み上げたときの度数の累積和

累積相対度数 (cumulative relative frequency)　相対度数の累積和

表 1.6　度数分布表

階級番号	階　級 (以上)～ (未満)	階級値 m_i	度　数 f_i	相対度数 f_i/n	累積度数 Σf_i	累積相 Σf
1	a_0～a_1	m_1	f_1	f_1/n	f_1	f_1
2	a_1～a_2	m_2	f_2	f_2/n	f_1+f_2	$(f_1+$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	a_{k-1}～a_k	m_k	f_k	f_k/n	$f_1+f_2+\cdots+f_k$	$(f_1+f_2+$
計			n	1		

2.1.2　ヒストグラム (histogram)

また**柱状グラフ、度数柱状図表**とも呼ばれる

連続型データ (continuous data) **の場合**　各階級を横軸として、柱の面積を度数と比例するように高さを定める

単峰型 (unimodal) **と双峰型** (bimodal)　峰が二つ以上ある分布（双峰型）の場合、層別化によって、峰がひとつの単純な分布（単峰型）が現れる

階級数の求め方　観測値の数を n、階級数を k とすると

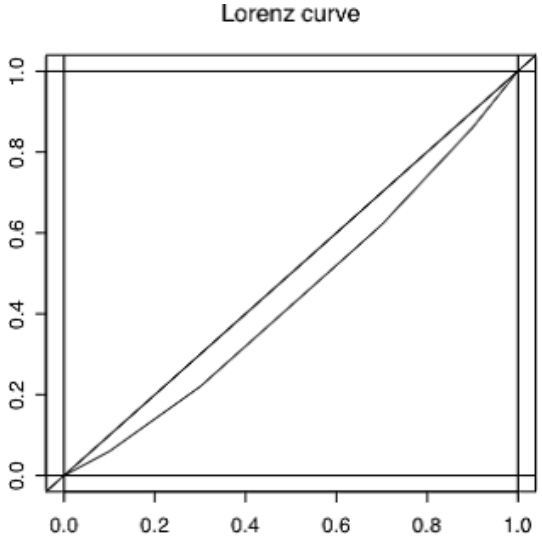
- 平方根則 $k=\sqrt{n}$
- スタージェスの公式 $k \doteqdot 1+log_2 n = 1+(log_{10}n)/(log_{10}2)$
- 他にまだいろいろな方法がある

階級幅の求め方　階級の幅を w、データの最大値 x_n と最小値 x_1 から 、
レンジ (range)(範囲) R は $R=x_n-x_1$ 、
そして、幅は $w=R/m$。

離散型データ (discrete data) **の場合**　（柱の横幅は一定として）高さが度数を反映するように描き、柱と柱の間に隙間を設ける（例：図 2.10）

2.1.3 ローレンツ曲線 (Lorenz curve)

累積相対度数を組み合わせて描いた折れ線である
1905 年にアメリカの経済学者マックス・ローレンツが発表した



2.1.4 測定の尺度

測定 (measurement) が依っている基準を 4 尺度 (scales) に分類した

名義（名目）尺度 (nominal scale)　ある個体（対象）が他とは異なるか同一かという判断のみの基準 (例：血液型)
この時、平均値を求めても意味がなく、最頻値は求められる

順序尺度 (ordinal scale)　ある個体が他より‘ 大きい ’、他より‘ 良い ’、他より (何かについて)‘ 多い ’といえる判断の基準 (例：授業評価)
これらの数値は大小関係にのみ意味がある

間隔尺度 (interval scale)　ある個体は他よりもある単位によって〜だけ多い (少ない) といえる判断の基準 (例：温度)
したがって、数値の差のみに意味がある

比尺度 (ratio scale)　ある個体は他よりもある単位によって〜倍だけ多い (少ない) といえる判断の基準 (例：体重)
比例尺度では数値の差と共に数値の比にも意味がある

2.2 代表値 (averages)

分布を代表する値、

2.2.1 平均 (mean)

平均には以下の 3 種類がある

算術平均 (arithmetic mean)　観測値 $x_1, x_2, \cdots, x_n$ の和をデータの大きさ n で割ったのものである

$$\bar{x} = \sum_{i=1}^{n} x_i/n = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{1}$$

観測値のとりうる値を $v_1, v_2, \cdots, v_k$、度数が $f_1, f_2, \cdots, f_k$ とすると、平均は

$$\bar{x} = \frac{f_1 v_1 + f_2 v_2 + \cdots + f_k v_k}{f_1 + f_2 + \cdots + f_k} \tag{2}$$

調和平均 (harmonic mean) x_H

$$\frac{1}{x_H} = \frac{1}{n}(\frac{1}{x_1} + \cdots + \frac{1}{x_n}) \quad \Longrightarrow x_H = \frac{n}{\sum\limits_{i=1}^{n} \frac{1}{x_i}} \tag{3}$$

したがって、調和平均値は、逆数の平均値の逆数である

幾何平均 (geometric mean) x_G

$$x_G = \sqrt[n]{x_1 \bullet x_2 \bullet \cdots \bullet x_n} \tag{4}$$

2.2.2 メディアン (中央値)(median)Me

データを大きさ順に並べたものを $x_{(1)} \leqq x_{(2)} \leqq \cdots \leqq x_{(n-1)} \leqq x_{(n)}$ とすると、

$$Me = \begin{cases} x_{(\frac{n+1}{2})} & (n = \text{奇数}) \\ \frac{1}{2}\{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\} & (n = \text{偶数}) \end{cases} \tag{5}$$

メディアンは、累積相対度数が 0.5 となるような値である

外れ値 (outlier)(**異常値**)　平均値とメディアンの値が近い場合は、概ねその値を中心として左右対称であることが多い。それに対してこの二つの値が離れているときは，対照性が崩れて右ないし左にゆがんでいるか，飛び離れた**外れ値**があることが多い。

四分位点 (quartile)　小さい順に並び替えられたデータを 4 等分したときの三つの分割点

- 第 1 四分位点　Q_1　25 % 分位点
- 第 2 四分位点　Q_2　50 % 分位点 (メディアン)
- 第 3 四分位点　Q_3　75 % 分位点

2.2.3　モード (最頻値)(mode)Mo

分布の峰に対応する値、度数分布表において度数が最大である階級の階級値である

2.3　散らばり (dispersion) の尺度

2.3.1　レンジ (範囲)(range)

データの最大値 x_n と最小値 x_1 から 、レンジ R は

$$R = x_n - x_1 \tag{6}$$

2.3.2　四分位偏差 (quartile deviation , semi-interquartile range)

データの第 3 四分位点 Q_3 と第 1 四分位点 Q_1 の隔たりの半分

$$Q = \frac{1}{2}(Q_3 - Q_1) \tag{7}$$

四分位偏差が大きいほど散らばった分布となる

2.3.3　平均偏差 (mean deviation)

各観測値が平均からどれだけ離れているかについての平均を求めたものである

$$d = \frac{1}{n}\{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \cdots + |x_n - \bar{x}|\} \tag{8}$$

2.3.4 分散と標準偏差

最小２乗法 (method of least squares) **と平均値**　n 個のデータ $x_1, x_2, \cdots, x_n$ を１個の値 a で代表させるとき、すなわち、様々な値をとっている n 個のデータをすべて a で置き換えるとき、偏差 (誤差)(deviation) が出てくる。ある特定のデータの誤差ではなく、全体としての誤差を小さくするには、誤差の和を小さくしたいということになる。絶対偏差 $f(a) = \sum_{i=1}^{n} \mid x_i - a \mid$、２乗偏差の和 $g(a) = \sum_{i=1}^{n} (x_i - a)^2$ を最小にする a を求め、その値をデータの代表値にする。

絶対偏差の和 $f(a)$ の最小値を求めることは難しい。２乗偏差の和 $g(a)$ を最小にするには、

$$\begin{aligned} g(a) &= \sum_{i=1}^{n} (x_i - a)^2 \\ &= \sum_{i=1}^{n} (x_i - \bar{x}) - (a - \bar{x})^2 \\ &= \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2\,(a - \bar{x}) \sum_{i=1}^{n} (x_i - \bar{x}) + \sum_{i=1}^{n} (a - \bar{x})^2 \\ &= \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(a - \bar{x})^2 \\ &\geqq \sum_{i=1}^{n} (x_i - \bar{x})^2 \end{aligned} \tag{9}$$

これより、$a = \bar{x}$ のとき、関数 $g(a)$ は最小値 $\sum_{i=1}^{n} (x_i - \bar{x})^2$ をとることがわかる。

分散 (variance)　偏差の２乗をとって、それの平均値である

$$S^2 = \frac{1}{n}\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\} \tag{10}$$

各階級の度数を f_i、階級値を v_i、m を階級数として

$$S^2 = \frac{f_1(v_1 - \bar{x})^2 + \cdots\cdots + f_k(v_n - \bar{x})^2}{f_1 + \cdots + f_k} \tag{11}$$

標準偏差 (standard deviation)

$$S = \sqrt{S^2} \tag{12}$$

2.3.5 変動係数 (coefficient of variation)

$$C.V. = S_x/\bar{x} \tag{13}$$

2.3.6　標準得点 (standard score)(標準化 (standardization))

１次変換 (linear transformation)　データを $z_i = ax_i + b$ のように、

$$\bar{z} = a\bar{x} + b \tag{14}$$

$$S_z{}^2 = a^2 S_x{}^2 \tag{15}$$

のように変わる
特に、$a = 1/S_x, b = -\bar{x}/S_x$ の場合、$\bar{z} = 0, S_z = 1$ になるので、この z をデータ x の**標準得点 (標準化)** という

　標準得点に一次変換 $T_i = S_z \cdot z_i + \bar{z}$ を施したものが**偏差値得点** (deviation value)
z_i, T_i はそれぞれ**Ｚ得点，Ｔ得点**と呼ばれることもある。

参考文献

[1] 橋本　智雄「入門統計学」　共立出版 (2004)

[2] 垂水　共之・飯塚　誠也「R/S-PLUS による統計解析入門」共立出版 (2006)