2014 年度 森林統計学 第 3 回平均値と標準偏差 配布資料-教科書の補足

1. “4.1 平均”と“4.2 標準偏差”(p.17～21)の計算方法

教科書の例では、「分類されたデータ」から計算する式・方法が紹介されている。度数分布表が作られているときには、この方法の方が、「分類されていないデータ」を用いる方法よりも簡単に平均や標準偏差を計算できる。しかし、「分類されたデータ」から計算すると、正確な値にならない（例えば教科書の例のように階級幅が 10 のとき、140 も 149 も「139.5～149.5」の階級に含まれるので「144.5」という値として計算されてしまうため)。

計算機やパソコンが気軽に使える現在では、「分類されていないデータ」が利用できるときには、「分類されていないデータ」を用いる方法で平均・標準偏差を算出するのが標準である。

※平均については p.18, 9～11 行に「各標本値をそれに対する階級値におきかえて求めた平均は，一般に元の測定値の平均とは多少異なる．それゆえ，公式(4)から求めた平均値は公式(3)による正しい平均値の近似値にすぎない.」、標準偏差については p.21, 7～8 行に「分類されていない元のデータが利用できるならば，平均の計算のときと同様に，標準偏差の計算は公式(10)より公式(9)によって求めねばならない.」とされている通り。

・教科書 p.19, 図 6 に示された分布のデータでの具体的な計算例

データ1 (教科書p.19, 図6 上の図の分布)

連番 i	値 X_i		$X_i - \overline{X}$	$\left(X_i - \overline{X}\right)^2$
1	X_1 =	4	−2.0	4.0
2	X_2 =	5	−1.0	1.0
3	X_3 =	6	0.0	0.0
4	X_4 =	7	1.0	1.0
5	X_5 =	8	2.0	4.0
(n = 5)	計	30		10.0
平均値	$\overline{x}$:	6.0		
	分散	s^2:		2.5
	標準偏差	s:		1.6

$\sum_{i=1}^{n} X_i$ データの合計

$\sum_{i=1}^{5} (X_i - \overline{X})^2$ 偏差平方和

$\frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \overline{X})^2$

データ2 (教科書p.19, 図6 下の図の分布)

連番 i	値 X_i		$X_i - \overline{X}$	$\left(X_i - \overline{X}\right)^2$
1	X_1 =	2	−4.0	16.0
2	X_2 =	4	−2.0	4.0
3	X_3 =	6	0.0	0.0
4	X_4 =	8	2.0	4.0
5	X_5 =	10	4.0	16.0
(n = 5)	計	30		40.0
平均値	$\overline{x}$:	6.0		
	分散	s^2:		10.0
	標準偏差	s:		3.2

データの合計: 30

偏差平方和: 40.0

※計算の手順は、**データの合計**を算出→データ数で割って**平均値**を算出→平均値を用いて**偏差平方和**を算出→偏差平方和を「データ数－1」で割って**分散**を算出→分散のルートをとって**標準偏差**を算出、となる。分散は元のデータを 2 乗した単位（元データが kg なら kg²）であることに注意。分散のルートをとって標準偏差にすると、元データと同じ単位になる（元データが kg なら kg)。

分散の定義式には、n（データ数）で割って計算する(5)式（分類されたデータに対しては(6)式）と、$n-1$ で割って計算する(7)式（分類されたデータに対しては(8)式）の 2 種類がある。対象とするデータが母集団の場合（あるいは母集団とみなされる場合）には前者を用いるが、一般にそれは特別な場合である（(5)式で計算される分散を特に「母分散」と呼ぶ)。通常のデータは標本とみなされる場合がほとんどであり、(7)式（分類されたデータに対しては(8)式）を用いる。いずれの場合も、分散の平方根が、標準偏差（母分散に対しては特に「母標準偏差」という）である。

2.“4.3 標準偏差の意味”(p.22～24)

どんなデータであるかを第 3 者に伝える必要があるとき、元データのすべてを示してこと細かに説明できる場合は少ない。多くの種類のデータを一度に説明しなければならないときには、ヒストグラムも使わずに、要点の統計数値のみしか示せないことはよくある。そのような時に最もよく使われる 2 つの統計数値が、平均値と標準偏差である。従って、逆の立場から考えると、平均値と標準偏差から、そのデータがおおむねどのような分布をしているのかを想像できなければならない。

標準偏差から元データの詳細な分布形状を知ることはできないが、分布の広がり方（ばらつき具合）を知ることはできる。そのことを実感するために、平均値±標準偏差（$\bar{x}\pm s$）の範囲内に全体の何割のデータが含まれているか、平均値±標準偏差の 2 倍（$\bar{x}\pm 2s$）の範囲内に全体の何割のデータが含まれているか、といったことを課題や練習問題の分布で確認しておくとよい。

教科書図 7 の例では、分類されたデータからこれらの値を求めている（階級内の個数は比例配分で求めている)。分類されていないデータが利用できる場合には、比例配分の計算をする必要はなく、データを大きい順に並べ替えた表などから数え上げればよい。

・標準偏差の比較

図 7 の例のように、同じ平均値の場合には標準偏差の大小でバラツキの大きさの違いを比較できる。平均値が大きく異なる場合（例えば平均値 5.0cm の小石の大きさのデータと平均値 53.8cm の岩石の大きさのデータに対して、標準偏差が前者で 3.0cm、後者で 10.0cm として後者の方がバラツキが大きいとは直ちには言えない）や測定値の単位自体が異なる場合（例：体重 kg の分布と身長 cm の分布）には、単純に標準偏差の値を比較することはできない。

そのような場合に使える指標として、平均値に対する標準偏差の割合を示す「変動係数」がある（Coefficient of Variance；C.V. $= s / \bar{x}$ ；×100 して%表示する場合もある)。変動係数は無次元数なので、単位が異なるデータ間の比較にも使用できる。

3.“5. その他の記述的測度”(p.24～26)

・最頻値（モード；mode）

離散型変数の場合には、教科書の定義にあるように「最大の度数をもつ測定値」として最頻値を求めることができる。連続型変数の場合には、度数分布表における最大度数をもつ階級を、「最頻階級」（値は階級値を用いる）と定義できる。平均・中央値とあわせて代表値のひとつ。

・中央値

測定値を大きさの順に並べたときちょうど中央にくる測定値の値として定義される。データの総数 n が奇数の場合、$(n+1)/2$ 番目の測定値が中央値となる（例：n=11 の場合には 6 番目の測定値)。n が偶数の場合は、$(n+1)/2$ は「～.5」となるのでその前後の順位の測定値の平均が中央値となる（例：n=10 の場合には 5 番目と 6 番目の測定値の平均値)。

・四分位数

四分位数の算出でも、中央値の場合のように n が奇数か偶数かによって $(n+1)/4$、$(n+1)\times 3/4$ 番目となる測定値あるいはその順位をはさむ 2 つの順位の測定値の平均で第 1・第 3 四分位数を求めることができる（第 2 四分位数＝中央値)。最小値を第 0 四分位数、最大値を第 5 四分位数として、5 つの四分位数の間の差を見ることで、分布の形状をおおまかに判断できる。

・四分位範囲

この範囲内にあるデータは全体の 50%となる。最小付近と最大付近の値を除外するので、異常値の影響を受けにくい。四分位範囲の半分を「四分位偏差」として用いる場合もある。