現代日本論演習／比較現代日本論研究演習 I「統計分析の基礎」

第 8 講　平均と分散

田中重人 (東北大学文学部准教授)

1　中間試験について

問 1–2 が各 6 点、問 3 が 8 点 (合計 20 点)

問 1(3) について： SPSS の recode では、複数の割り当て規則が該当する場合、**前のほうが優先** で処理される。このため、`(lowest thru 50 = 1)(50 thru highest = 2)` と書いても `(lowest thru 50 = 1)(51 thru highest = 2)` と書いても結果は同じになる。しかし `(50 thru highest = 2)(lowest thru 50 = 1)` はちがう結果になる。

2　代表値と散布度

教科書 pp. 42–52 を読んで、「中央値」「四分位偏差」の計算方法を理解しよう。

3　平均値と標準偏差

平均 (mean):　総和をデータ数で割ったもの
分散 (variance):　平均値からの偏差の 2 乗値の平均
標準偏差 (standard deviation):　分散の平方根 (SD と書くことが多い)

教科書の表 2-1 (p. 48) で何が計算されているかを理解する

- 平均と標準偏差はセットで使う
- 尺度水準による制限

4　宿題

教科書 p. 52 の練習問題 2-3 について、平均値と標準偏差を計算せよ。計算の途中経過がわかるように解答すること。ISTU で来週水曜正午まで。

5　度数分布表のオプション

度数分布表の「統計量」オプションで「平均値」と「標準偏差」をチェック。

- 「記述統計」→「記述統計」でも出力できる。
- SPSS などの統計ソフトは、すこしちがう計算式で「標準偏差」を計算している (教科書 p. 48 注 6)。データが大きくなれば (およそ 200 以上なら) このことによるちがいはほとんどなくなるが、小さいデータ (たとえば 10 人程度) では大きなちがいになるので注意。

練習問題：「生活全般満足度」について、平均値と標準偏差を出力してみよう。

6　順序尺度の変数の「平均値」

平均値は、本来は、間隔尺度以上の水準の変数にしか使えない。しかし、実際には、一定条件を満たせば、順序尺度についても平均値をとっていいとする基準が使われている。

- 潜在的には間隔尺度のはず
- 測定のポイントが一定間隔

具体的には、4点以上の尺度であって、正規分布に近似している場合 (教科書 p. 53–59)。これは、「偶然の積み重ねで形成されるものは正規分布にしたがう」という仮定による。

「正規分布に近似」しているかどうかは、通常、つぎの３点で判断する。

- 単峰性
- 左右対称性 (歪度)
- 中央への集中度 (尖度)

SPSS でヒストグラムを描いて検討するとよい。

「度数分布表」の「統計量」オプションで「歪度」「尖度」を指定すると、正規分布との乖離度を統計的に検討できる。これらの値は、正規分布のとき０をとり、絶対値が大きくなるほど、正規分布から外れる。およそ ± 2 の範囲を超えていれば、正規分布からのずれが無視できない。

これらの条件を満たさない場合は非線形変換 (教科書 p.142–144) をおこなったり、順位に変換したりすることがある。あるいは、平均値を使わずに中央値を使って分析することもある。

なお、2 値の変数は、この条件にかかわらず間隔尺度とみなしてよいが、一定以上のデータ数があり、あまり偏っていないことが必要。

7　平均値の欠点

平均値は「はずれ値」(outlier) の影響を受けやすい。あまりにかけはなれたケースがあるときは

- 上下数%を取りのぞく (調整平均：教科書 p. 46)
- 順位に変換したり中央値を使って分析

などの方法を使うことがある。

また、極端なはずれ値がなくとも、左右非対称の分布の変数 (所得、人口、めったに起こらない現象の経験回数など) では、平均値より中央値の方が適切な代表値であることが多い。