第 2 章 脱線 点相関係数 point correlation coefficient

浅川 伸一

2006 年 10 月 26 日

課題の状況からニーズがありそうなので脱線する。

前期 χ^2 統計量、特にクロス表の独立性の検定で言及したとおり、クロス表から計算された χ^2 値が以下のように計算された。

0.1 分割表 (クロス表) の独立性の検定 (再録)

実測値と理論値の乖離に基づく独立性の検定の例を計算方法を示す。2×2 の分割表の場合、総データ数を n とすれば、ある属性が p_1 が起こる期待値は (2 項分布より) np_1 で与えられる。同様に、もう一方の属性 p_2 が起こる場合の期待値は np_2 で与えられる。もし属性 1 が得られたとしても属性 2 が無関係であるならば np_1 の値を知ったとしても np_2 の値は独立である、という。この場合、2 の表の値の中の一つの観測値を知ると他のすべての値が定まる。

表 1: 2×2 のクロス表、理論値の求め方

	p_1	$1-p_1$	合計
p_2	np_1p_2	$n(1-p_1)p_2$	np_2
$1-p_2$	$np_1(1-p_2)$	$n(1-p_1)(1-p_2)$	$n(1-p_2)$
合計	np_1	$n(1-p_1)$	n

得られた χ^2 値を周辺度数が固定されていると考えた場合の χ^2 の値の最大値で割った値の平方根をカテゴリーデータの相関と呼ぶことがある。

特に 2×2 のクロス表についての値を点相関係数 point correlation coefficient またはΦ (ファイ) 相関係数 phi correlation coefficient と呼ぶ。この相関係数は今までの相関係数 (他の相関係数と区別するときにはピアソンの積率相関係数 Peason's product moment correlation coefficient) と異なり $0 \leq \phi \leq 1$ の範囲となる。

クロス表の各セルの値を とすれば

	0	1
0	a	b
1	c	d

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(b+c)(b+d)}} \tag{1}$$

なる。

この ϕ は、カテゴリーに属する場合を 1 そうでない場合を 0 とみなしてピア損の積率相関係数を計算した値と等しくなる。

1 操作例

Excel 上での操作は煩雑であるが手で集計するよりはミスが少いだろう。以下の手順はエクセルのバージョンによっても異なるので一般的ではないが、

1. ウィンドウの最上行にあるメニューから [データ], [ピボットテーブルレポート] と選択してピボットテーブルウィザードを起動する
2. ウィザードの画面に従いデータ範囲等を設定する
3. ピボットテーブルメニューバーから変数の名前を行、列それぞれへドラッグアンドドロップする。
4. どちらかの変数を表の中央へもドラッグアンドドロップする。
5. 表の左上を右クリックし、フィールドの設定を選ぶ。
6. 集計の方法をデータの個数に変更する。

以上の操作によりクロス表ができる。他の手順によっても同じ数値を得ることができるだろう。