第3回　クロス表とχ二乗検定

１．クロス集計

クロス表とは2つの質的変数を組み合わせたカテゴリーの度数分布表。クロス表によって2つの質的変数間の関連を見ることができる。

クロス表の一般的な表現

Y / X	1	2	…	*j*	…	*J*	計
1	n_{11}	n_{12}	…	n_{1j}	…	n_{1J}	$n_{1\cdot}$
2	n_{21}	n_{22}	…	n_{2j}	…	n_{2J}	$n_{2\cdot}$
・	・	・		・		・	・
・	・	・		・		・	・
i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{iJ}	$n_{i\cdot}$
・	・	・		・		・	・
・	・	・		・		・	・
I	n_{I1}	n_{I2}	…	n_{Ij}	…	n_{IJ}	$n_{I\cdot}$
計	$n_{\cdot 1}$	$n_{\cdot 2}$		$n_{\cdot j}$		$n_{\cdot J}$	n

２．クロス表の例（１）

ここで使っているデータは、2000 年に東京都 5 区市で実施した質問紙調査（郵送法）によって得られたデータである。居住地点によって学歴構成が異なるかどうかを調べてみる。

chiten　居住地域 と q43　学歴 のクロス表

			q43　学歴				
			1　中学卒	2　高校卒	3　短大卒	4　大学卒	合計
chiten 居住地域	1　港区	度数	17	49	29	83	178
		chiten　居住地域 の %	9.6%	27.5%	16.3%	46.6%	100.0%
	2　大田区	度数	14	79	26	59	178
		chiten　居住地域 の %	7.9%	44.4%	14.6%	33.1%	100.0%
	3　世田谷区	度数	8	50	30	88	176
		chiten　居住地域 の %	4.5%	28.4%	17.0%	50.0%	100.0%
	4　清瀬市	度数	22	92	37	57	208
		chiten　居住地域 の %	10.6%	44.2%	17.8%	27.4%	100.0%
	5　あきる野市	度数	40	95	37	48	220
		chiten　居住地域 の %	18.2%	43.2%	16.8%	21.8%	100.0%
合計		度数	101	365	159	335	960
		chiten　居住地域 の %	10.5%	38.0%	16.6%	34.9%	100.0%

港区の回答者の 46.6 %は大学卒、世田谷区の回答者の 50.0 %も大学卒。
清瀬市の回答者の 10.6 %は中学卒、あきる野市の回答者の 18.2 %は中学卒。

このことから、地域によって、住民の教育水準が異なることが推測できる。
　→居住地と住民の学歴とのあいだには関連がある。

*この場合、因果関係は複雑。世田谷区に住むと学歴が高くなると言えるかどうか。むしろ､学歴の高い人が世田谷区や港区に住むようになると考えるほうが真実に近いであろう。

３．クロス表の例（２）
　クロス表による因果関係の分析
　理論的に因果関係の方向がはっきりしている場合がある。例：人種とガンによる死亡率（ザイゼル『数字で語る』より引用)。分析目的によって、パーセンテージを縦にとるか横にとるかが決まるので注意。

	ガン	その他	計
白人	139,627	1,055,804	1,195,432
黒人	9,182	169,391	178,573
計	148,809	1,225,195	1,374,004

	ガン	その他	計
白人	93.8	86.2	87.0
黒人	6.2	13.8	13.0
計	100.0	100.0	100.0

（タテのパーセント）

	ガン	その他	計
白人	11.7	88.3	100.0
黒人	5.1	94.9	100.0
	10.7	89.3	100.0

（ヨコのパーセント）

原則：独立変数を表側におけば、ヨコに 100 %になるように計算し、タテに比較する。
　　表頭におけば、タテに 100 %になるように計算し、ヨコに比較する。
ただし、どちらが原因でどちらが結果であるかは、理論的な仮説の問題で、クロス表そのものは関連を示すだけである。（**因果関係**と**相関関係**）

４．クロス表の関連の度合いを見る――χ 二乗検定とクラメールの V 係数

　比較する比率が､まったく一致することは稀である｡どのくらい違いがあればよいのか？
　1)悉皆調査の場合、１％でも違えば、厳密に１％違うというのが事実。
　　→しかし、関連の強さを示すことはできないだろうか。→クラメールの V 係数
　2)標本調査の場合、標本誤差を考慮する必要がある。どのくらい違えば、母集団においても違いがあるといえるのだろうか。→χ^2 検定

χ^2検定（カイ二乗検定）
　I × J のクロス表において、２つの変数 X と Y に関連があるといえるかどうかを検定する。
　かりに X と Y との間に関連がなかった場合（X と Y が独立）に、このクロス表がどう

なるか（期待度数）を考え、それと現実のクロス表とのずれに注目する。ずれが大きければ大きいほど、独立でない、つまり関連がある可能性が高くなる。

$$\chi_0^2=\sum_i\sum_j\frac{(n_{ij}-F_{ij})^2}{F_{ij}}$$

ここで、$F_{ij}=\frac{n_{i\cdot}\,n_{\cdot j}}{n}$

（周辺度数によって比例配分した期待度数）

XとYが独立の場合、このχ_0^2の標本分布は、自由度$(I-1)(J-1)$のχ^2分布に近似することが知られている。

説明のための架空の例

表１　（観測されたデータ）

	従属変数Y		
独立変数X	1	2	計
1	9	31	40
2	31	29	60
計	40	60	100

もし、ＸとＹとの間に全く関連がなければ、各セルは表２のようになるはず（太字のところは変わらない）。

表２　（各セル：完全に独立の場合の期待値 Fij）

	従属変数Y		
独立変数X	1	2	計
1	16	24	**40**
2	24	36	**60**
計	**40**	**60**	**100**

表１と表２の各セルの差を計算すると、表３のようになる。

表３（各セル：観測値－期待値）

	従属変数Y		
独立変数X	1	2	計
1	-7	+7	
2	+7	-7	
計			

次に、各セルの差を２乗する。

表４　各セル：（観測値－期待値）^{2}

	従属変数Y		
独立変数X	1	2	
1	49	49	
2	49	49	

これを各セルの期待値（表２）で割る。

表５　各セル：（観測値－期待値）^{2}／期待値

	従属変数Y		
独立変数X	1	2	
1	3.06	2.04	
2	2.04	1.36	

各セルの値を総計したものがカイ自乗値

$\chi^2 = \Sigma\Sigma$　（観測値－期待値）^{2}／期待値

＝ 3.06+2.04+2.04+1.36

＝ 8.5

この数字をカイ自乗分布表にあてはめる。

自由度（I-1)(J-1)、 この場合は２×２表だから（２－１）×（２－１）＝１

自由度は１。よって、χ二乗分布表より、「危険率５％以下で独立性を棄却できる」。

つまり、関連があると言える。

クラメールのV（クラマーのV、クラメアのV）

$$V=\sqrt{\frac{\chi^2}{(m\text{-}1)n}}$$　ただし、*m*はIとJの小さい方

$$=\sqrt{\frac{8.5}{(2-1)\times 100}}$$

$=0.29$

※クラメールのＶは、最小値が０で、最大値が１となる。

※χ二乗値は、最小値は０（独立）だが、最大値は不定。