社会調査入門/社会調査論

第7章　質的データの統計分析

立命館大学経済学部

寺　脇　　拓

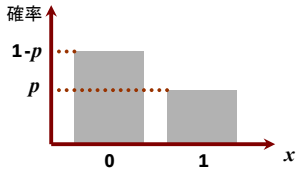
本章の概要

本章では、クロス集計表を用いた独立性の検定を中心に質的データの統計分析方法を学ぶ[1)]。

1. 比率の推定と検定

1.1 比率の推定

- **ベルヌーイ分布**(Bernoulli distribution)
  - 浄水器の所有率を推定したいとする。
  - 浄水器の所有の有無を表す変数をxで表し、「浄水器をもっている」を1、「浄水器をもっていない」を0で表す。
  - 母集団の浄水器を持っている人の割合をpで表すとすると、その母集団から無作為抽出されたxの値は確率変数となり、それは次のようなベルヌーイ分布に従う。



  - ベルヌーイ確率変数は、0か1の値をとり、1をとる確率がp、0をとる確率1-pの離散確率変数である。
  - ベルヌーイ確率変数の平均はp、分散は$p(1-p)$となる。

- 点推定
  - 母集団のタコ焼き器の所有率pの自然な推定量は、xの標本平均であろう。

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - この $\bar{x}$ は，pの不偏かつ一致推定量であり、その意味で望ましい推定量である。
- 区間推定
  - xの平均はp、分散は$p(1\text{-}p)$ であるので，定理5.3(中心極限定理)より，nが大きいとき，次のzは標準正規分布に従う.

$$z = \frac{\bar{x} - p}{\sqrt{p(1-p)}/\sqrt{n}}$$

  - 従って、次の式が成立する。

$$P\left(\bar{x} - 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq p < \bar{x} + 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = 0.95$$

  - nが大きいときには、$\sqrt{p(1-p)}$ を $\sqrt{\bar{x}(1-\bar{x})}$ に置き換えることで、母比率(xの母平均)pの95%信頼区間を次のように計算することができる。

$$\left(\bar{x} - 1.96\frac{\sqrt{\bar{x}(1-\bar{x})}}{\sqrt{n}},\ \bar{x} + 1.96\frac{\sqrt{\bar{x}(1-\bar{x})}}{\sqrt{n}}\right)$$

  - 90%信頼区間は、上記の1.96を1.65に置き換えたもので表され、99%信頼区間は、それを2.58に置き換たもので表される。

1.2 比率の検定

- 母比率が50%を超えているかどうかを検定する。

$$H_0:\ p = 0.5$$
$$H_1:\ p > 0.5$$

- 帰無仮説が正しいとき、そしてnが十分に大きいとき、次のzは標準正規分布に従う。

$$z = \frac{\bar{x} - 0.5}{0.5/\sqrt{n}}$$

- 片側検定なので、もし観測値から計算されるzの値が1.65を超えるならば、95%水準で帰無仮説は棄却される。
- 一般に、母比率がαを超えているかどうかを検定するときには、次のzを用いて片側検定を行う。

$$z = \frac{\bar{x} - \alpha}{\sqrt{\alpha(1-\alpha)}/\sqrt{n}}$$

■ 母比率の差の検定

定理7.1

母数p_aのベルヌーイ母集団から無作為抽出された大きさn_aの標本の標本平均を $\bar{x}_a$ 、母数p_bのベルヌーイ母集団から無作為抽出された大きさn_bの標本の標本平均を $\bar{x}_b$ で表すとき、nが大きければ、その標本平均の差 $\bar{x}_b - \bar{x}_a$ は、以下の平均m 、分散vの正規分布に近似的に従う。

$$m = p_b - p_a$$
$$v = \frac{p_a(1-p_a)}{n_a} + \frac{p_b(1-p_b)}{n_b}$$

- 両比率に差があるかどうかだけを検定したい場合には、帰無仮説、対立仮説は次のように表される。

$$H_0:\ p_b - p_a = 0$$
$$H_1:\ p_b - p_a \neq 0$$

- 帰無仮説が正しいとき、次のzは、nが大きいときに標準正規分布に従う。

$$z = \frac{\bar{x}_b - \bar{x}_a}{\sqrt{p_0(1-p_0)\left(\frac{1}{n_a}+\frac{1}{n_b}\right)}}$$

  - ただし、p_0=p_a=p_bである。
- p_0の値はわからないので、これを次のp_0^*で置き換えて、標準正規分布に基づく検定を行う。

$$p_0^* = \frac{n_a\bar{x}_a + n_b\bar{x}_b}{n_a + n_b}$$

2. クロス集計結果の統計分析

2.1 独立性の検定

- 居住地(都市部か農村部か)と里山保全に対する評価との間に何らかの関係があるかどうかを調べたいとしよう[2)]。
- 両質問の回答形式が単一回答で、選択肢が二つのとき、クロス集計表は次のように表される。
  - 表側の行数がr、表頭の列数がkのクロス集計表を**$r \times k$クロス集計表**という。この場合は、2×2クロス集計表ということになる。

表7.1 居住地別に見た里山保全に対する評価

	都市部	農村部	全体
里山保全は重要	a	c	a+c
里山保全は重要でない	b	d	b+d
合計	a+b	c+d	n=a+b+c+d

- もし居住地と里山保全に対する評価との間に何の関係もないならば、すなわち

居住地が都市部か農村部かということと里山保全を重要だと考えるかどうかということが互いに独立である

ならば、次の四つの式が成立するはずである。

帰無仮説

$$\frac{a}{n} = \frac{a+b}{n}\times\frac{a+c}{n}, \quad \frac{b}{n} = \frac{a+b}{n}\times\frac{b+d}{n},$$
$$\frac{c}{n} = \frac{c+d}{n}\times\frac{a+c}{n}, \quad \frac{d}{n} = \frac{c+d}{n}\times\frac{b+d}{n}$$

- そして、期待される各セルの観測値数は次のようになる。

$$\bar{a} = \frac{(a+b)(a+c)}{n}, \quad \bar{b} = \frac{(a+b)(b+d)}{n},$$
$$\bar{c} = \frac{(c+d)(a+c)}{n}, \quad \bar{d} = \frac{(c+d)(b+d)}{n}$$

- このとき、次の式で計算される(検定)統計量は、自由度1のカイ二乗分布に近似的に従う。

$$\chi^2 = \frac{(a-\bar{a})^2}{\bar{a}} + \frac{(b-\bar{b})^2}{\bar{b}} + \frac{(c-\bar{c})^2}{\bar{c}} + \frac{(d-\bar{d})^2}{\bar{d}}$$

  - ここで、「バー」がついていないものは実際の観測値数を、「バー」がついているものは期待される観測値数を表している。
- 観測値からこの値を計算し、カイ二乗分布表に基づいて、検定を行う。
- これを**独立性の検定**という。
- 有意水準に対応する棄却域は次のとおり。
  - 10%： $\chi^2 > 2.71$
  - 5%： $\chi^2 > 3.84$
  - 1%： $\chi^2 > 6.64$

- この検定の対立仮説は、「居住地が都市部か農村部かということと里山保全を重要だと考えるかどうかということは互いに独立ではない」ということになるが、それはここでは、「都市部の人には里山保全を重要だと思う人が多い」か「農村部の人には里山保全を重要だと思う人が多い」のいずれかを意味する。
- もし、都市部で里山保全が重要だと思う人の割合が、農村部のそれを上回っているのであれば、対立仮説を前者に、逆であれば、対立仮説を後者にしてしまってよい。

■ 補正カイ二乗検定

- 2×2クロス集計表で、a、b、c、dのいずれかが5以下の場合には、検定統計量がカイ二乗分布でうまく近似されない。
- このときには、検定統計量を次のように補正する(**イエーツの補正**)。

$$\chi^2 = \frac{(|a-\bar{a}|-0.5)^2}{\bar{a}} + \frac{(|b-\bar{b}|-0.5)^2}{\bar{b}} + \frac{(|c-\bar{c}|-0.5)^2}{\bar{c}} + \frac{(|d-\bar{d}|-0.5)^2}{\bar{d}}$$

2.2 オッズ比

- 「居住地が都市部か農村部か」と「里山保全を重要だと思うかどうか」について、母集団の構成が次の二つのケースを考える。

表7.2 居住地別に見た里山保全に対する評価（母集団の構成、ケースA）

	都市部		農村部		全体	
	度数	%	度数	%	度数	%
里山保全は重要	9000	**42.9%**	2000	**25.0%**	11000	37.9%
里山保全は重要でない	12000	57.1%	6000	75.0%	18000	62.1%
合計	21000	100.0%	8000	100.0%	29000	100.0%

表7.3 居住地別に見た里山保全に対する評価（母集団の構成、ケースB）

	都市部		農村部		全体	
	度数	%	度数	%	度数	%
里山保全は重要	14000	**66.7%**	2000	**25.0%**	16000	55.2%
里山保全は重要でない	7000	33.3%	6000	75.0%	13000	44.8%
合計	21000	100.0%	8000	100.0%	29000	100.0%

- どちらも「都市部の人には里山保全が重要だと思う人が多い」ということになるが、ケースAでは「少しだけ多い」のに対して、ケースBでは「とても多い」ことを示している。

- この関係の強さを測る指標の一つに**オッズ比**(Odds Ratio)がある。

表7.4 居住地別に見た里山保全に対する評価（母集団の構成、一般表記）

	都市部	農村部	全体
里山保全は重要	α	γ	$\alpha+\gamma$
里山保全は重要でない	β	δ	$\beta+\delta$
合計	$\alpha+\beta$	$\gamma+\delta$	$\Omega=\alpha+\beta+\gamma+\delta$

$$Odds\ Ratio = \frac{\frac{\alpha}{\alpha+\beta}}{\frac{\beta}{\alpha+\beta}} \Big/ \frac{\frac{\gamma}{\gamma+\delta}}{\frac{\delta}{\gamma+\delta}} = \frac{\alpha\delta}{\beta\gamma}$$

  - オッズ比は二つの質問の回答が独立であるとき1となり、関係が強いほど1から離れる。
- また、オッズ比の自然対数をとった**対数オッズ比**もしばしば示される。
  - 対数オッズ比は、二つの質問が独立であるとき0となり、関係が強いほど0から離れる。

■ オッズ比の推定

- 点推定
  - 得られたクロス集計結果から$\alpha\delta/\beta\gamma$、すなわちad/bcを計算する。
- 区間推定
  - オッズ比をφとすると、標本対数オッズ比は、平均が$\ln\varphi$、分散が次のvの正規分布で近似される。

$$v = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

  - 従って、標本対数オッズ比の95%信頼区間は、点推定される標本オッズ比 $\hat{\phi}$ (=ad/bc)を用いて、次のように表される。

$$\left(\ln\hat{\phi} - 1.96\sqrt{v},\ \ln\hat{\phi} + 1.96\sqrt{v}\right)$$

  - オッズ比の形に戻すと、その95%信頼区間は次のようになる。

$$\left(\hat{\phi}\exp(-1.96\sqrt{v}),\ \hat{\phi}\exp(1.96\sqrt{v})\right)$$

    - expは指数関数(e≒2.72を底とする数式のべき乗)を表している。

2.3 一般的な独立性の検定

表7.5 情報提供の有無別に見た遺伝子組み換え食品の安全性評価

	情報あり	情報なし	全体
安全だと思う	a	d	a+d
危険だと思う	b	e	b+e
わからない	c	f	c+f
合計	a+b+c	d+e+f	n=a+b+c+d+e+f

- r×kのケースにおいても、基本的に検定の手続きは同じ。
- 上記のような3×2の場合、2×2のケースと同様に、まずは各セルの期待される観測値数を計算する。

$$\bar{a} = \frac{(a+b+c)(a+d)}{n}, \quad \bar{b} = \frac{(a+b+c)(b+e)}{n},$$
$$\bar{c} = \frac{(a+b+c)(c+f)}{n}, \quad \bar{d} = \frac{(d+e+f)(a+d)}{n},$$
$$\bar{e} = \frac{(d+e+f)(b+e)}{n}, \quad \bar{f} = \frac{(d+e+f)(c+f)}{n}$$

- 次に検定統計量を次のように計算する.

$$\chi^2 = \frac{(a-\bar{a})^2}{\bar{a}} + \frac{(b-\bar{b})^2}{\bar{b}} + \frac{(c-\bar{c})^2}{\bar{c}} + \frac{(d-\bar{d})^2}{\bar{d}} + \frac{(e-\bar{e})^2}{\bar{e}} + \frac{(f-\bar{f})^2}{\bar{f}}$$

- r×kの場合、検定統計量は**自由度(r-1)×(k-1)のカイ二乗分布**に従う。
- この例では、(3-1)×(2-1)で、自由度は2となる。
- あとはカイ二乗分布表に基づいて検定を行う。

■ 注

1. 本章は、岩田(1983)第7章、第8章、森棟(2000)の第6章、第8章を参考にした。
2. 農林統計では、各市町村は「都市的地域」、「平地農業地域」、「中間農業地域」、「山間農業地域」に区分される。ここでの「都市部」は「都市的地域」を、「農村部」は残りの三つの地域をイメージしている。各地域の定義については、http://www.maff.go.jp/yougo_syu/toukei.htmlを参照のこと。

■ 引用文献

- 岩田暁一(1983)『経済分析のための統計的方法 第2版』、東洋経済。
- 森棟公夫(2000)『統計学入門 第2版』、新世社。