

カテゴリカルデータ分析入門

宮田 庸一

平成 26 年 1 月 11 日

目次

第 1 章 質的（カテゴリカル）データに対する統計処理	3
1.1 カテゴリカルデータ	3
1.2 クロス集計表	3
1.3 仮説検定	4
1.4 カイ二乗検定	5
1.5 p 値 (p-value)	7
1.6 関連性の強さ	7
1.7 カイ二乗検定の応用	8
1.7.1 経営への応用	8
1.7.2 食中毒の原因特定	8
1.8 一般のクロス集計表	9
1.9 期待度数 (授業では説明しないかも)	9
1.10 章末問題	10
付 録 A R による統計処理	12

データ分析入門を担当される先生方へ：本講座の担当していただきありがとうございます。データ分析入門は、経済学部に入学者 1,2 年生が受けるべき講座として考えております。すでにご存じかと思われそうですが、本学の新入生は、数学のレベルに大きな差があり、数学のできない学生にとっては統計学はできれば避けたい科目の一つとなっています。しかし結局卒業論文を書く段階になって、統計処理ができないために右往左往している現状があります。このためこの講座では、統計解析の数学的な仕組みの詳細な説明は行わず、以下の点に重点を置いて授業を進めていただきたいと思います。

1. 統計解析により出力されたグラフ、結果の見方、統計手法の使い方を中心に説明する。統計処理の仕組みはイメージだけで十分。
2. 必要とする予備知識は、四則演算 ($+$ $-$ \times \div) および $\sqrt{\quad}$ とする。このため確率論は極力使用しない。微分積分は使用しない。
3. 統計処理は統計ソフトで処理する。
4. 統計解析の理論的な仕組みを知りたい学生は、統計学 I, 統計学 II を受講するよう勧めてください。

そこですが、グラフ、平均、分散などの記述統計はよいとしても、仮説検定のような説明に確率論を用いないでどのように説明するのかわかるかもしれません。このため、クロス集計表、仮説検定の指導案のイメージを作成しましたので、参考にいただければと思います。しかしこれはあくまでも**授業の一案**ということです。このため上記の重点項目に従っていれば、この指導案に従わず、ご担当の先生の説明のしやすい方法で授業をしていただいてもかまいません。よろしくお願いします。

第1章 質的（カテゴリカル）データに対する統計処理

✓ 今回の講義で使用するファイル:カイ二乗検定.xlsx

1.1 カテゴリカルデータ

我々が考えるデータは気温や重さのように数えられるものに限られません。この章では人の満足度や好みのように、数えることのできないデータに対する統計手法について説明を行う。ちなみに数えることのできないを**質的データ**、もしくは**カテゴリカルデータ**という。さて以下のようなアンケートを150人にとったとしよう。

質問1	あなたの性別は。 0. 女 1. 男
質問2	あなたはタバコを吸いますか？ 0. 吸う 1. 吸わない

このようなアンケートをとると以下のようなデータが得られる。尚、質問1に対する回答を項目1、質問2に対する回答を項目2とおいた。

名前	項目1	項目2
A	0	1
B	1	1
C	0	0
⋮	⋮	⋮
Z	0	1

表 1.1:

1.2 クロス集計表

次に項目1と項目2に対してどのような傾向があるのか調べるために、以下のような表にまとめる。この表のことを**分割表**もしくは**クロス集計表**という。特にこの場合、質問1の項目が2つ、質問2の項目が2つあるので、正確には2×2分割表という。また表1.2の61や39などの数値が入っている場所(枠)のことを**セル**という。例えば男であり、かつ喫煙が有る場所は、1行目2列目の場所にあるので、(1,2)セルという。

項目 1 \ 項目 2	項目 2		計
	無	有	
男	61	39	100
女	8	42	50
計	69	81	150

表 1.2:

1.3 仮説検定

表 1.2 の分割表を眺めてみると、なんとなく男性と女性ではタバコの吸う人の割合が異なるように見える。しかし男性と女性では調査した人数が異なる上に、たまたま得られたデータが表 1 のようになったのかもしれない。このため、性別と喫煙の有無が独立であるかどうかを調べるためには、仮説検定と呼ばれる手法を行う必要がある。仮説検定を行うためには、**帰無仮説** (Null hypothesis) と**対立仮説** (Alternative hypothesis) と呼ばれる 2 つの仮説を設定する必要がある。帰無仮説は通常 H_0 という記号で表し、対立仮説は H_a という記号で表す。上記の分割表 1.2 においては、下記のように定める。

H_0 : 質問 1 に対する解答と質問 2 に対する解答には関係がない

H_a : 質問 1 に対する解答と質問 2 に対する解答には関係がある

帰無仮説の方に”関係がない”と決める必要があるが、とりあえずはそういうものだと考えてほしい。仮説検定とは何かを大ざっぱにいうと、帰無仮説 H_0 と対立仮説 H_a どちらが正しいかを与えられたデータから判断するということである。ただし厳密には、”対立仮説 H_a が正しい”か、もしくは”対立仮説 H_a が正しいという証拠がない”かをデータから判断することである。これについては後で説明する。ここで”対立仮説が正しい”と結論づけることを、帰無仮説 H_0 を**棄却** (Reject) するといい、”対立仮説 H_a が正しいという証拠がない”と結論づけることを、帰無仮説 H_0 を**採択** (Accept) するという。仮説検定は帰無仮説を棄却するか採択するかを判断するわけなのだが、100%正しい結論はあり得ないという理念を持っている。この大ざっぱなイメージを説明するために以下の例を考えよう。

判決 \ 真実	真実	
	無罪	有罪
有罪	第 1 種の誤り	○
無罪	○	第 2 種の誤り

決定 \ H_0	H_0	
	正しい	間違い
H_0 を棄却	第 1 種の誤り	○
H_0 を採択	○	第 2 種の誤り

裁判の例においては、被告が本当は有罪なのに無罪を言い渡される誤りと、本当は無罪であるのに有罪を言い渡される 2 種類の誤りがある。実際の仮説検定においても、帰無仮説を採択、もしくは棄却するという 2 つの選択肢があるため、2 種類の誤りがある。帰無仮説 H_0 が正しいのに、 H_0 を棄却してしまう誤りのことを**第 1 種の誤り (Type I error)** であるといい、一方で帰無仮説が間違っているのに、帰無仮説を採択してしまう誤りのことを**第 2 種の誤り (Type II error)** という。裁判の例において、被告が本当は有罪なのに無罪となる確率を小さくしようとすると、(直接証拠ではなく) 状況証拠などを積極的に採用しなくてはならない。しかしそのような行動をすると、被告が本当は無罪なのに有罪となる確率を大きくすることになってしまう。このことは、第 1 種の誤りの確率と第 2 種の誤りの確率は同時に小さくすることはできないことが直感的に

わかる。仮説検定においても同じことがいえて、完全に正しい判断を下すことはできない。このため帰無仮説 H_0 が正しいのに、誤って H_0 を棄却してしまう確率を何パーセント（通常5%）認めることにする。この確率は**有意水準**という。このため、統計学を用いて100%正しい結論を下すことはできないということを忘れないでほしい。

最後に仮説検定をどのように行うかのイメージを説明する。まず帰無仮説 H_0 が正しいと仮定する。ⁱ もし帰無仮説 H_0 が正しいときに得られる結果に反する状況が現れたときには、帰無仮説 H_0 を棄却するというわけである。逆に帰無仮説 H_0 に反する状況が現れないときには、帰無仮説 H_0 を採択するということになる。

例えば、殺人罪の容疑で A 氏が逮捕されたとし、容疑者 A が「私はやっていない」といっているとしよう。ここで“容疑者 A は無罪である”と仮定する。このときに、防犯カメラなどで容疑者 A の犯行が映っているときには、これは“容疑者 A は無罪である”とした仮定に反する結果であるため、“容疑者 A は無罪である”という主張は棄却されることになる。しかし“容疑者 A は無罪である”とした仮定に反する結果がないからといって、容疑者 A は無罪であるといえるか？ 犯行の状況がばれていないだけかもしれない。これは容疑者 A は有罪である根拠がないという言い方が適切であることが直感的にわかる。同様の理由から、帰無仮説が採択されるということは、帰無仮説を**積極的に**支持するのではなく、対立仮説が正しいという証拠がないということからの帰無仮説に対する**消極的な**支持であることがわかる。

1.4 カイ二乗検定

仮説検定を行うためには、検定統計量と呼ばれるものを計算する必要がある。ここでは表 1.2 を以下の一般的な形で表すことを考える。

項目 1 \ 項目 2	項目 2		計
	B_1	B_2	
A_1	f_{11}	f_{12}	$f_{11} + f_{12}$
A_2	f_{21}	f_{22}	$f_{21} + f_{22}$
計	$f_{11} + f_{21}$	$f_{12} + f_{22}$	n

表 1.3:

また帰無仮説、対立仮説を以下のように定める。

H_0 : 項目 1 と項目 2 には関係がない

H_a : 項目 1 と項目 2 には関係がある

ここで $f_{.1} = f_{11} + f_{21}$, $f_{.2} = f_{12} + f_{22}$, $f_{1.} = f_{11} + f_{12}$, $f_{2.} = f_{21} + f_{22}$ とする。このとき、カイ二乗統計量 χ^2 を以下のように定義するⁱⁱ。

$$\chi^2 = \frac{\left(f_{11} - \frac{f_{1.}f_{.1}}{n}\right)^2}{\frac{f_{1.}f_{.1}}{n}} + \frac{\left(f_{12} - \frac{f_{1.}f_{.2}}{n}\right)^2}{\frac{f_{1.}f_{.2}}{n}} + \frac{\left(f_{21} - \frac{f_{2.}f_{.1}}{n}\right)^2}{\frac{f_{2.}f_{.1}}{n}} + \frac{\left(f_{22} - \frac{f_{2.}f_{.2}}{n}\right)^2}{\frac{f_{2.}f_{.2}}{n}} \quad (1.1)$$

$$= \frac{n(f_{11}f_{22} - f_{12}f_{21})^2}{(f_{11} + f_{12})(f_{11} + f_{21})(f_{12} + f_{22})(f_{21} + f_{22})} \quad (1.2)$$

ここで (1.1) 式の意味は今の時点では考えなくてよい。(1.1) 式は 2×2 分割表だけではなく 3×3 , 4×3 の分割表のように、より一般的な分割表における仮説検定を行う際に

ⁱこれは帰無仮説が本当は正しいかどうかなどは考えなくてよい。形式的に帰無仮説 H_0 が正しいと仮定すればよい。

ⁱⁱ χ はギリシャ文字で“カイ”という。

必要となる. 実際に手で計算する際には, (1.2) 式を用いる. これも式の詳しい意味は今の時点では考えなくてよい. χ^2 は得られたデータが帰無仮説 H_0 からどのくらいいずれがあるのかを示している. このため χ^2 が大きな値を取ることは, H_0 を棄却できることを意味している. これを体感するために以下の問 1.1 の 2 つの問題を解いてほしい. しかし χ^2 がどのくらい大きな値をとればよいのかについては, どこかで境界線を引かなくてはならない. 実際には以下のようにして仮説検定を行う.

有意水準 5% の仮説検定 (境界値を用いた方法)

2 × 2 分割表において

$$\begin{cases} \chi^2 > \chi_{0.05}^2(1) & \iff \text{帰無仮説 } H_0 \text{ を棄却.} \\ \chi^2 \leq \chi_{0.05}^2(1) & \iff \text{帰無仮説 } H_0 \text{ を採択.} \end{cases}$$

$\chi_{0.05}^2(1)$ は境界値 (きょうかいち) と呼ばれる値で自由度 1 のカイ二乗分布における上側 5% 点となる. Excel では,

$$=\text{CHISQ.INV.RT}(0.05, 1)$$

と入力すればよい.

【例】表 1.2 のデータに対して

$$\chi^2 = \frac{150 \cdot (61 \cdot 42 - 89 \cdot 8)^2}{100 \cdot 69 \cdot 81 \cdot 50} = 18.37091. \quad (1.3)$$

となる. (配布された Excel ファイルカイ二乗検定.xlsx では自動的に計算してくれる) Excel で=CHISQ.INV.RT(0.05, 1) と入力して計算すると $\chi_{0.05}^2(1) \div 3.84$ となる. よって $18.37091 > 3.84$ より, 帰無仮説 H_0 は棄却される. つまり性別差がタバコを吸うか否かに影響を与えているという結論になる.

問 1.1 (1) "1 枚のコインを 2 回投げる" という実験を 100 回繰り返したところ以下の結果が得られたⁱⁱⁱ. このとき, χ^2 を求めよ.

		2 回目		
		表	裏	計
1 回目	表	32	25	57
	裏	19	24	43
計		51	49	100

(2) 袋の中に「当」と書かれたカードが 3 枚, 「は」と書かれたカードが 2 枚入っているとす. この袋の中から"1 枚ずつ計 2 枚のカードを無作為に取り出す. ただし 1 回目に取り出したカードは元に戻さない" という実験を 100 回繰り返したところ以下の結果が得られた. 直感的には, 1 回目の結果と 2 回目の結果は関係 (関連) がありそうである. このとき χ^2 を求めよ.

		2 回目		
		当	は	計
1 回目	当	26	31	57
	は	32	11	43
計		58	42	100

ⁱⁱⁱ筆者は実際に実験をしてみた結果である. 直感的には, 1 回目の結果と 2 回目の結果は独立であることがわかる.

Cramer(クラメル)の連関係数

$$V = \sqrt{\frac{\chi^2}{n(\min(r, c) - 1)}}$$

ここで r は行の数, c は列の数, $\min(r, c)$ は行の数と列の数で小さい方の数を表す.
また χ^2 はカイ二乗統計量とする.

Cramerの連関係数 V は0以上1以下の値を取り, 関連性が弱い場合は0に近い値をとり, 関連性が強い場合は1に近い値をとる. Cramerの連関係数は 2×2 分割表だけでなく, 一般の $r \times c$ 分割表に対して適用することができる. 項目の関連性の強さを表す指標は他にもあるが, 興味のある方は藤井 (2010, p.53), 宿久, 村上, 原 (2011, pp.56-57) を参考にしてほしい. 尚, 配布された Excel ファイルカイ二乗検定.xlsx においては, Cramerの連関係数 V を自動的に計算してくれる.

注意 2×2 分割表 (即ち $r = 2, c = 2$) において, Cramerの連関係数は $V = \sqrt{\chi^2/n}$ となる. 実は表 1.1 のデータを対のデータ $(0, 1), (1, 1), (0, 0), \dots, (0, 1)$ とみなしたときの相関係数を r とするとき, $\chi^2/n = r^2$ となることが知られている. これより Cramerの連関係数 V が項目1と項目2の関連性の強さを表していることが直感的にわかるであろう.

1.7 カイ二乗検定の応用

1.7.1 経営への応用

ここでは, クロス集計表, カイ二乗検定がどのように役立つのかを説明する. まずは西内先生の著書「統計学が最強の学問である」の例 (p.82-85) を紹介しよう. ある企業が商品の新しいデザインを考えて, それが購入に結び付くのかを調査するために10万人に対して調査したとする.

	商品を購入	商品を購入せず	計
既存のデザイン	9500	90500	100000
新しいデザイン	9600	90400	100000
計	19100	180900	200000

ここで既存のデザインの場合, 10万人の中で購入したのが9500人なので, 9.5%が購入したということになる. 一方で新しいデザインは9.6%が購入したことを意味する. わずか0.1%という数値を侮ってはいけない. もし100億円の売り上げがある企業であれば, 0.1%売上げが上昇するだけでも1億円の売り上げ増ということになる. しかし0.1%の購入した割合の上昇が, たまたまなのか, もしくは意味のあるものかを仮説検定で検証するということになる. 実際, この場合 p 値 = 0.447 より帰無仮説は採択されるので, デザインにより売上に差があるとはいえない状況になる.

1.7.2 食中毒の原因特定

公衆衛生学においてよく知られた方法であるが, 食中毒をおこしたときに原因となる食べ物を調査に基づいて特定するための道具として, しばしばカイ二乗検定を行う. 西岡, 他 (2006) によると, 有症者の共通食は当該施設の食事に限られ, 12月16日, 医師

から食中毒届が提出されたことから、同日、当該施設で提供された食事を原因とする食中毒と断定したと書かれている。さらに有症者便から *S. Enteritidis* 及び *C. jejuni* を検出したことから、病因物質は *S. Enteritidis* 及び *C. jejuni* と断定したとある。ここでは地鶏（ささみ）の刺身を食べたかどうかを症状が発症した施設において調査した西岡、他（2006）のデータを用いる。

症状の有無 \ 摂取の有無	摂取の有無		計
	食べた	食べない	
有症者	71	29	100
健康者	25	36	61
計	96	65	161

ここで p 値 = 0.000166 < 0.05 より帰無仮説は棄却されるので、食中毒を発生した原因の一つとして地鶏（ささみ）の刺身があると結論づけられる。このような調査を多くの食品に対して適用することで、食中毒の原因となった食べ物を特定するのに役立てることができる。

1.8 一般のクロス集計表

これまでは 2×2 集計表についての、仮説検定を説明してきた。しかしこれは一般の $r \times c$ 分割表に対する仮説検定を行うことができる。ここでは詳しくは述べないが、配布された Excel ファイル「カイ二乗検定.xlsx」において、 p 値を見れば、仮説検定を行うことができる。

1.9 期待度数 (授業では説明しないかも)

統計ソフトを用いて仮説検定を行う上では表 (おもて) には出てこないが、内部では期待度数と呼ばれる値を計算している。ここでは期待度数、検定統計量について説明を行う。まず表 1.2 を思いだしてほしい。ここで男の数、女の数などそれぞれの合計を全体の 150 で割ると以下のようなになる。

項目 1 \ 項目 2	項目 2		計
	無 (B_1)	有 (B_2)	
男 (A_1)			$\frac{100}{150}$
女 (A_2)			$\frac{50}{150}$
計	$\frac{69}{150}$	$\frac{81}{150}$	1

表 1.4:

ここで A_1, A_2, B_1, B_2 はそれぞれ事象を表す記号として用いる。

このとき、 $E_{11} = 150 \cdot \frac{100}{150} \cdot \frac{69}{150}$ を 1 行 1 列目における期待度数という。期待度数を説明する前に、高校 1 年生に勉強した独立という概念を思い出そう。

事象の独立

$P(A \cap B) = P(A)P(B)$ が成り立つとき、2 つの事象 A, B は独立であるという。

この時、前の章で説明した帰無仮説 H_0 、対立仮説 H_a は厳密には、以下のように表すことができる。

H_0 : 全ての $i = 1, 2, j = 1, 2$ に対して事象 A_i と B_j は独立^{iv}.

H_a : 独立でない事象 A_i と B_j が存在する.

もし帰無仮説が正しい場合には, A_1 と B_1 は独立なので, $P(A_1 \cap B_1) = P(A_1)P(B_1) = \frac{100}{150} \cdot \frac{69}{150}$ となる. このとき, A_1 かつ B_1 となる度数は, $150 \cdot \frac{100}{150} \cdot \frac{69}{150}$ と推定できる. 一言でいうと, 期待度数とは, 帰無仮説が正しいと仮定したとき予測できる値ということになる. 同様に, E_{12}, E_{21}, E_{22} も期待度数を考えればよい. さてカイ二乗統計量を思い出そう. これは以下の形で表すことができる.

$$\chi^2 = \frac{(f_{11} - E_{11})^2}{E_{11}} + \frac{(f_{12} - E_{12})^2}{E_{12}} + \frac{(f_{21} - E_{21})^2}{E_{21}} + \frac{(f_{22} - E_{22})^2}{E_{22}} \quad (1.4)$$

ここで $\frac{(f_{11} - E_{11})^2}{E_{11}}$ の項に着目しよう. f_{11} は表 1.3 において A_1 かつ B_1 である度数の数を意味する. 一方で期待度数 E_{11} は, 帰無仮説 H_0 が正しいと仮定したときに予測される A_1 かつ B_1 の度数である. このため $\frac{(f_{11} - E_{11})^2}{E_{11}}$ は, 実際に観測された A_1 かつ B_1 の度数と帰無仮説が正しいと仮定したもとの予測される度数のずれを測っているものと考えることができる. 同様に, A_1 かつ B_2 の場合, A_2 かつ B_1 の場合, A_2 かつ B_2 の場合においても, 実際に観測された度数と帰無仮説が正しいと仮定したもとの予測される度数のずれを $\frac{(f_{ij} - E_{ij})^2}{E_{ij}}$ の形で表すことができる. このため, 式 (1.4) は, 実際に得られたデータが帰無仮説からどのくらいずれているのかを測っていることがわかる.

細かい話は省略するが, 帰無仮説 H_0 の仮定した下で, χ^2 は自由度 1 のカイ二乗分布に従うことが知られている.

1.10 章末問題

1. 以下のクロス集計表は, T 大学の学生と W 大学の学生に対して正規分布における確率を求める問題に対する正答もしくは間違った数を示している.

大学名	解答		
	正解	不正解	計
T 大学	84	62	146
W 大学	65	37	102
計	149	99	248

このとき, この問題に対して T 大学か W 大学かで正答するか否かに影響を与えているかを調べたい. 有意水準 5% で仮説検定を行うこと.

^{iv}つまり A_1 と B_1 は独立, A_1 と B_2 は独立, A_2 と B_1 は独立, A_2 と B_2 は独立ということ

関連図書

- [1] 藤井 良宜 (2010) カテゴリカルデータ解析, 共立出版.
統計ソフト R を用いて質的データに対する統計処理を説明してある.
- [2] データ分析入門, 第 7 版, 慶應 SFC データ分析教育グループ編.
JMP という統計ソフトを用いることで, 数学的な予備知識を四則演算 (+ − × ÷) にとどめた形で, 統計手法を説明してある. ただし, 標本平均 (つまり平均値) の説明すらソフトまかせにしているので, 自習書としてはつらい. しかしデータ分析で重要なことが書いてあるので, 授業で使う教科書という意味ではよいと思う.
- [3] 西内 啓 (2013) 統計学が最強の学問である, ダイヤモンド社.
- [4] 西岡 達彦, 小野 高明, 井上 圭祐, 佐藤 千絵, 岩井 徹, 中島 敏浩, 藤井 千香子, 重東 和宏, 三宅 伸子, 箱田 泰二, 岡 敬 (2006) 飲食店における Salmonella Enteritidis 及び Campylobacter jejuni による食中毒事件について. URL: <http://www.hiro-vet.or.jp/learning/pdf/magazine-h1812/P077-079.pdf>
- [5] 宿久 洋, 村上 亨, 原 恭彦 (2011) 確率と統計の基礎 I, ミネルヴァ書房
ここで説明した統計手法の厳密な説明が書いてある。数学科出身の学生にとっては, このような本の方がよいのかもしれない。

付録A Rによる統計処理

この章は宮田個人的なメモです。無視してください。

表 1.2 の分割表を作成し、カイ二乗検定を行う場合:

```
example22<-matrix(c(61,39,8,42),ncol=2,byrow=T) #2 × 2 分割表の作成
chisq.test(example22,correct=FALSE)
```

```
example23<-matrix( c(279,73,225,165,47,191), nrow=2, byrow=T) #2 × 3 分
割表の作成
chisq.test(example23,correct=FALSE)
```

```
example44<-matrix(c(67,128,20,5,3,22,10,5,155,180,40,20,9,49,10,10),nrow=4,byrow=T)
#4 × 4 分割表の作成
chisq.test(example44,correct=FALSE)
```

クラメールの V を計算するためのコマンド

カイ二乗値を計算する関数

```
chisq <- function(mat){ex<- outer(rowSums(mat), colSums(mat))/sum(mat) # 期
待値 sum((mat-ex)^2/ex) # カイ二乗値
}
```

φ 係数を求める関数 (chisq 関数が必要)

```
phi <- function(mat){sqrt(chisq(mat)/sum(mat))}
```

コンティンジェンシー係数を求める関数 (chisq 関数が必要)

```
contingency <- function(mat){
  temp <- chisq(mat)
  sqrt(temp/(sum(mat)+temp))}
}
```

クラメール係数を求める関数 (phi, chisq 関数が必要)

```
cramer <- function(mat){phi(mat)/sqrt(min(nrow(mat), ncol(mat))-1)}
```