第11章
テキストにおける
差異の分析と特徴抽出

佐々木研究室
荒井悠有

1.概要

-この章では、度数データを用いてカイ2乗統計量による差異の分析、テキストにおける特徴語を抽出する方法について説明する。

2.適合度の検定

カイ2乗値
-実測値と理論値の乖離を測る統計量

	項目１	項目２	・・・	項目c	横合計
実測値	n_1	n_2	・・・	n_c	N
理論値(確率)	p_1	p_2	・・・	p_c	1
期待度数	Np_1	Np_2	・・・	Np_c	N

$$\chi^2 = \sum_{i=1}^{c} \frac{\left(n_i - Np_i\right)^2}{Np_i}$$

・例

-6面体のサイコロのゆがみについてテストを行う

面	1	2	3	4	5	6	合計
結果	13	15	28	26	28	10	120
理論値（確率）	1/6	1/6	1/6	1/6	1/6	1/6	1
期待度数	20	20	20	20	20	20	120

$$\chi^2 = \frac{(13-20)^2}{20} + \frac{(15-20)^2}{20} + \cdots + \frac{(10-20)^2}{20}$$

$$= 16.9$$

このカイ2乗値は自由度5のカイ2乗分布に従う。

適合度の検定では、実測度数と理論度数が同じであると仮説を立てる。

このサイコロの例では・・・

帰無仮説：サイコロの目の出方に差がない

対立仮説：サイコロの目の出方に差がある

カイ2乗分布表より、自由度5の有意水準0.05のカイ２乗値は11.07→16.9よりも小さい

この場合、帰無仮説は棄却される。

→有意水準0.05で目の出方に差がないとはいえない。

参考 - カイ2乗分布表

df	0.05	0.01
1	3.841	6.635
2	5.991	9.21
3	7.815	11.34
4	9.488	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21

適合度の検定

-データが予測される確率の通りになっているかどうかを判定する。

判定基準

・$X^2 \leqq X^2_\alpha$ → 帰無仮説を棄却できない

・$X^2 > X^2_\alpha$ → 帰無仮説を棄却する　　＊X^2_α：有意水準αでのX^2の値

ちなみに、求めたカイ2乗値の上側の確率(p値)と有意水準αを
比較して判定することも可能

・$p \geqq \alpha$ → 帰無仮説を棄却できない

・$p < \alpha$ → 帰無仮説を棄却する

3. 分割表の検定(独立性の検定)

分割表の行（あるいは列）のパターンが同じであるか否かを、カイ2乗分布に基づいて統計的に判断する方法
→期待度数の乖離の度合いを測り、その度合いを用いて判断する。

$$\text{期待度数} = \frac{\text{横の合計} \times \text{縦の合計}}{\text{総合計}}$$

芥川龍之介の読点のデータ

	大正15年	昭和2年	横の合計
は、	41	82	123
が、	103	429	532
て、	26	6	32
ら、	107	303	410
に、	30	61	91
その他	290	803	1093
縦の合計	597	1684	2281

期待度数

	大正15年	昭和2年	横の合計
は、	32.2	90.8	123
が、	139.2	392.8	532
て、	8.4	23.6	32
ら、	107.3	302.7	410
に、	23.8	67.2	91
その他	286.1	806.9	1093
縦の合計	597	1684	2281

分割表におけるi行j列の度数をn_{ij}、対応する期待度数をE_{ij}とすると

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij}-E_{ij}\right)^2}{E_{ij}}$$

行×列

このカイ2乗値は自由度(r-1)(c-1)のカイ2乗分布に従う。

$$\chi^2 = \sum_{i=1}^{6}\sum_{j=1}^{2}\frac{\left(n_{ij}-E_{ij}\right)^2}{E_{ij}}$$

$$= \frac{(41-32.2)^2}{32.2} + \frac{(82-90.8)^2}{90.8} + \cdots + \frac{(803-806.9)^2}{806.9}$$

$$\fallingdotseq 68.6$$

このカイ2乗値は、自由度(6-1)(2-1)=5のカイ2乗分布に従う。

分割表の検定では、行・列のパターンが同じであると仮説を立てる。

帰無仮説：行（あるいは列）のパターンに差がない
対立仮説：行（あるいは列）のパターンに差がある

有意水準を0.05とすると自由度5のカイ2乗値は11.07
得られたカイ2乗値(68.6)はこの値より大きいので帰無仮説は棄却される。

3.1 イェーツ補正

カイ2乗値を用いて2×2の分割表（データ数が少ない）を検定する場合は、式を補正する必要がある。

2×2の分割表

	b_2	b_3	横の合計
a_1	n_{11}	n_{12}	n_{1+}
a_2	n_{21}	n_{22}	n_{2+}
縦の合計	n_{+1}	n_{+2}	n_{++}

2×2の分割表のカイ2乗値は

$$\chi^2 = \frac{n_{++}(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}}$$

イェーツは、よりカイ2乗分布に近似するように、
次のように補正を施して用いる方法を提唱した。

$$\chi^2{}_Y = \frac{n_{++}(|n_{11}n_{22} - n_{12}n_{21}| - 0.5n_{++})^2}{n_{1+}n_{2+}n_{+1}n_{+2}}$$

3.2 フィッシャーの正確確率検定

一般的に、分割表の期待度数が5以下のセルが全体の25%以上であるときには、カイ2乗検定は不適切であるとされている。
→フィッシャーは超幾何分布を用いて、2×2の分割表の正確確率検定の統計量を導出した。

$$p(n_{11}) = \frac{{}_{n_{1+}}C_{n_{11}} \times {}_{n_{2+}}C_{n_{21}}}{{}_{n_{++}}C_{+1}}$$

$$= \frac{n_{1+}!\,n_{2+}!\,n_{+1}!\,n_{+2}!}{n_{11}!\,n_{12}!\,n_{21}!\,n_{22}!\,n_{++}!}$$

2×2の分割表においては、周辺度数が固定されたとき、1つのセルの値が決まれば、その他の3つのセルの値は確定される。
このように求めたp値（p(x)、x＝0,1,2,･･･,n_{1+}）を用いた検定をフィッシャーの正確確率検定（直接確率検定）とよぶ。

例.

紅茶実験データ

実際	識別結果 ミルク	紅茶	合計
ミルク	3	1	4
紅茶	1	3	4
合計	4	4	8

「紅茶にミルクを注ぐ」と「ミルクに紅茶を注ぐ」という異なる2つの方法で入れた紅茶を用意し、その味から紅茶の入れ方を推定するテスト。

p(x),x=0,1,2,3,4の値は、次のようになる。

$$p(0) = \frac{{}_4C_0 \times {}_4C_4}{{}_8C_4} = \frac{4! \times 4! \times 4! \times 4!}{0! \times 4! \times 4! \times 0! \times 8!} = 0.0143$$

p(1)=0.2286, p(2)=0.5143, p(3)=0.2286, p(4)=0.0143

両側検定はp(0)+p(1)+p(3)+p(4)
片側検定はp(0)+p(1)（左側検定）、あるいはp(3)+p(4)（右側検定）

3.3 尤度比統計量とその他

分割表のカイ2乗値の近似値として、尤度比がある。

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \log\left(\frac{n_{ij}}{E_{ij}}\right)$$

分割表の独立検定には、カイ2乗統計量が多く用いられるが、モデルの比較や選択には、尤度比統計量が多く用いられている。

分割表に関しては、検定を行う以外に、相関係数のような関連の度合を示す量で分析する場合もある。分割表の行と列の項目の関連の度合を測るためのカイ2乗値を用いた指標として、ファイ連関係数φ、ピアソンの連関係数C、クラメールの連関係数Vがあり、次のように定義される。式中のkは行と列の少ない方をとる。

$$\phi = \sqrt{\frac{\chi^2}{n_{++}}} \qquad C = \sqrt{\frac{\chi^2}{\chi^2 + n_{++}}} \qquad V = \sqrt{\frac{\chi^2}{n_{++}(k-1)}}$$

これらの値が大きいほど、分割表の行と列の関連性が強いと判断する。

カイ2乗値を用いて、検定を行う際に注意しなければいけないのは、結果が総度数に依存することである。
→総度数が大きいほど仮説が棄却されやすい。
総度数を50以上にするべきでると言われているが、その上限に関する目安はない。

4. カイ2乗値を用いた特徴語の抽出

例. 2つのグループのテキストにおける特徴語の検出

-芥川龍之介の読点のデータの特徴語の抽出

項目ごとのカイ2乗値とp値

	大正15年	昭和2年	カイ2乗値	p値
は、	41	82	3.069	0.08
が、	103	429	16.205	0
て、	26	6	48.102	0
ら、	107	303	0.001	0.981
に、	30	61	1.913	0.169
その他	290	803	0.107	0.743

項目「て、」のカイ2乗値が最も大きい。
→両テキストにおける使用頻度の差が最も顕著に現れている。

例. 2つ以上のグループにおける特徴語の抽出

-安倍、福田、麻生元総理の3つの所信表明演説分に用いられた特徴語の抽出

ここではテキストに用いられている名詞のみについて分析する。

3つのテキストに用いられた名詞

	安倍	福田	麻生
こと	27	27	23
国民	15	16	22
ため	26	13	13
よう	13	21	6
国	29	4	4
日本	23	5	9
もの	3	10	16
我が国	11	9	6
わたし	0	0	24
⋮	⋮	⋮	⋮
合計	1450	1116	964

語によっては、テキスト中では全く使用されない場合もあるので通常のカイ2乗値を求める式では正しい情報が得られない恐れがある。

→フィッシャーの正確確率を求めて用いる。

3つのテキストの名詞のp値

	安倍	福田	麻生	フィッシャーのp値
わたし	0	0	24	2.3×10^{-14}
民主党	0	0	12	1.6×10^{-7}
国	29	4	4	2.7×10^{-5}
もの	3	10	16	0.00033
不安	0	4	9	0.00034
立場	0	8	1	0.00047
行政	0	9	4	0.00082
安心	2	12	2	0.00117
財政再建	1	0	6	0.00251
将来	2	8	0	0.00264
私	13	8	0	0.00351
何	0	1	5	0.00426
国民生活	0	5	0	0.00465
⋮	⋮	⋮	⋮	⋮

この表で、最も特徴のある語は「わたし」

「私、わたし、わたくし」のp値と順位

	安倍	福田	麻生	フィッシャーのp値
民主党	0	0	12	1.6×10^{-7}
国	29	4	4	2.7×10^{-5}
もの	3	10	16	0.00033
不安	0	4	9	0.00034
私、わたし、わたくし	13	8	25	0.00045
立場	0	8	1	0.00048
行政	0	9	4	0.00082
安心	2	12	2	0.00117
⋮	⋮	⋮	⋮	⋮

「わたし」という語が「民主党」ほど麻生の所信表明演説分の特徴語にはならない。

→テキストマイニングを行う時は、同義語の処理が非常に重要