Data Mining Theory

DM-03: 独立性の検定に関する補足

さまざまな検定まとめ(3)

6. 2x2クロス集計表 → 2軸に有意な関連はある？
   - Fisherの正確確率検定
   - 度数が小さいとき、偏りが大きいときに、χ2乗検定より正確
   - odds, p_val = ss.fisher_exact(DataFrame)
7. 一般のクロス集計表→ 2軸に有意な関連はある？
   - χ^2 検定 (独立性の検定)
   - 各度数は5以上であることが望ましい
   - chi2, p_val, dof, expected = ss. chi2_contingency(DataFrame)

   クロス集計表
8. 観測度数(x_obs)と理論度数(x_exp) → 両者は有意に異なる？
   - χ^2 検定 (適合度の検定)
   - 各度数は5以上であることが望ましい
   - chi2, p_val = ss. chisquare(x_obs, f_exp=x_exp)

   「一様」かどうかの検定なら
   f_exp=は省略可

dm-03-assign1

- dm-04フォルダのshop_sales.csvは、ある店の都道府県(Prefs)ごとの一部店舗をサンプルとして、店舗種類(フランチャイズ / 直営)と、ある商品の販売数(キャンペーン前(Sales1)と後(Sales2)の2データ)をまとめたものである。以下の検定を行うノートブックを作成し、結果と ipynb/html ファイルを提出せよ。いずれも有意水準は5%とする。

1. 全データを用いたとき、キャンペーン後のSales2の方が、キャンペーン前のSales1よりも販売数が増加したといえるか。
2. TokyoとSaitamaのSales2の販売数は、有意に異なるといえるか。
3. 都道府県と店舗種類のクロス集計表を作成し、都道府県と店舗種類の間に関連があるかどうかを判断せよ (カイ2乗検定を用いる)
4. (発展) 各都道府県の店舗数の比率は、都道府県の人口 (Mを100万人として、Tokyo:14.0M, Saitama:7.3M, Chiba:6.3M, Kanagawa:9.2M) の比率と一致しているとみなしてよいか。


```
ct = pd.crosstab(df['Pref'], df['Type'])
display(ct)
```

Type Pref	direct	franchise
Chiba	13	15
Kanagawa	28	30
Saitama	24	25
Tokyo	34	31

```
chi2, p, dof, expected = ss.chi2_contingency(ct)
print(chi2, p, dof, expected)
```

```
0.35072743456882416 0.9502219446727902 3 [[13.86   14.14 ]
 [28.71   29.29 ]
 [24.255 24.745]
 [32.175 32.825]]
```

expected

Type Pref	direct	franchise
Chiba	13.86	14.14
Kanagawa	28.71	29.29
Saitama	24.255	24.745
Tokyo	32.175	32.825

Pref列とType列が独立だとしたときに期待される表

expectedの求め方

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.contingency.expected_freq.html#scipy.stats.contingency.expected_freq

```
ct = pd.crosstab(df['Pref']
display(ct)
```

Type Pref	direct	franchise
Chiba	13	15
Kanagawa	28	30
Saitama	24	25
Tokyo	34	31
	99	101

Type列がPref列と独立なら、
Prefの値に関係なく
direct/franchiseの総数に比例
した配分になるはず

99：101に比例配分すると

28 → 13.86, 14.14
58 → 28.71, 29.29
49 → 24.255, 24.745
65 → 32.175, 32.825

Type Pref	direct	franchis e
Chiba	13.86	14.14
Kanagawa	28.71	29.29
Saitama	24.255	24.745
Tokyo	32.175	32.825

expectedの補足:行と列が独立だと仮定し、行和列の和から要素を推計

PrefとTypeが独立 → P(Pref ∩ Type)=P(Pref)P(Type)

✓ 例えばP(Chiba ∩ direct)=P(Chiba)P(direct)
P(Chiba), P(direct)はそれぞれ列和、行和から計算し、

$$P(\text{Chiba}) = 28/200, P(\text{direct}) = 99/200$$

表の要素はカウント数だから、確率*総カウント数で
(28/200)*(99/200)*200=28*99/200=13.86

一般にi行j列の要素はi行の和をT_i, j列の和をT_j, 総数をNとして、

$$\frac{T_i}{N} \times \frac{T_j}{N} \times N = \frac{T_i T_j}{N}$$

Type	direct	franchise	All
Pref			
Chiba	???		28
Kanagawa		?	58
Saitama			49
Tokyo			65
All	99	101	200

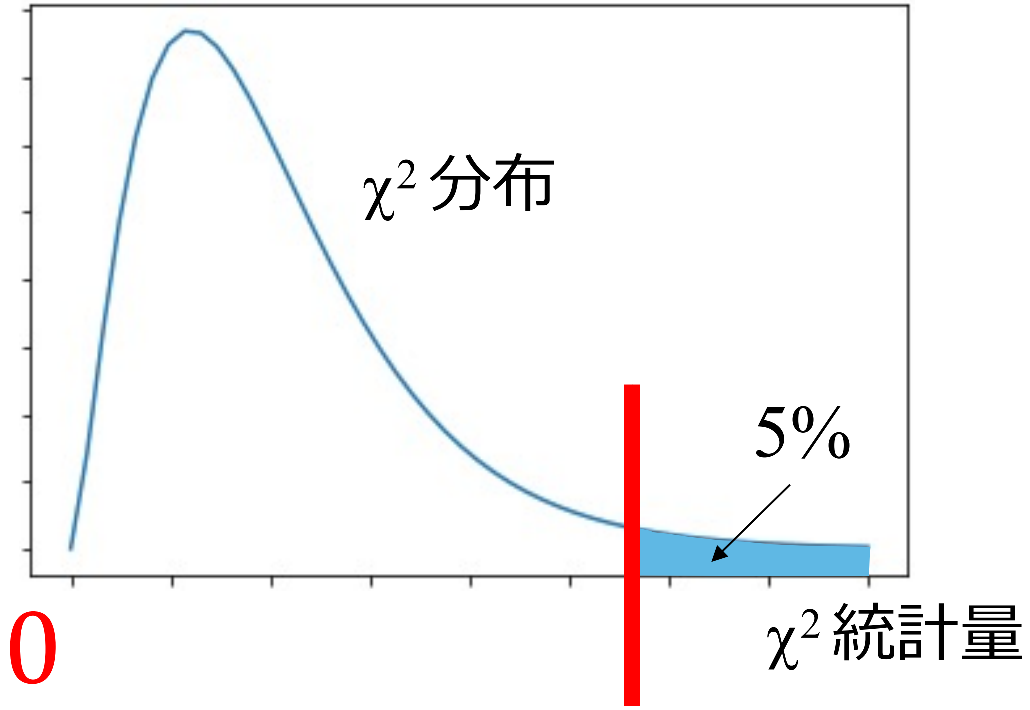
Type	direct	franchise	All
Pref			
Chiba	13.860	14.140	28
Kanagawa	28.710	29.290	58
Saitama	24.255	24.745	49
Tokyo	32.175	32.825	65
All	99	101	200

http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_HypothesisTesting-ChiSquare/BS704_HypothesisTesting-ChiSquare_print.html

検定

chi2, p_val, dof, expected = ss. chi2_contingency(DataFrame)

- Pref列とType列が独立なら、実際のクロス集計表 (4行2列) と expected表からのずれから計算できる「chi2 (χ^2) 統計量」は、自由度 (4-1) * (2-1) = 3 のカイ2乗分布にしたがうことが知られている
  - χ^2 統計量: 表の各セルについて、expected表の値と実際のクロス集計表の値の差の2乗を実際のクロス集計表の数値で割ったものを、表のすべてのセルについて足し合わせた量



t (Welch) 検定と同様に、χ^2 統計量が、χ^2 分布における有意水準5%の臨界値 (赤線) よりも大きくなった場合は、H0 (Pref列とType列は独立) を棄却して、H1 (Pref列とType列は関連がある)を採用する。