統計的仮説検定　その4

1　カイ2乗検定

1.1　χ^2 分布

これまで何度も統計量としての平方和 $S = \sum(x_i - \bar{x})^2$ の計算を行ってきた．この S は標本のばらつきを表す指標となるが，S の分布は，当然ながら標本の大きさ n と母分散 σ^2 と関係する．そこで，平方和 S と母分散 σ^2 との比を考えると，無次元量となり，種々の比較の分布として使用可能となる．その分布の形は S の分布と同じになるので，S に代わって S/σ^2 の分布を考えていく．このとき，

$$\chi^2 = \frac{S}{\sigma^2}$$

のように表した分布を自由度 $\phi = n - 1$ の χ^2 分布と呼ぶ．図1に $\phi = 3,\ 4,\ 5$ および6のときの χ^2 分布の例を示す．この分布に従う統計量の検定を行うために，χ^2 分布表が作成されている．配布資料を参照してほしい．

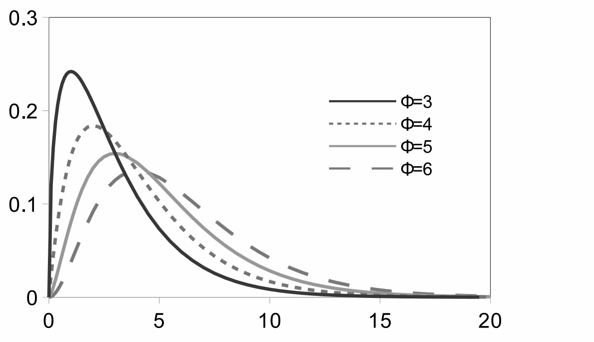

図 1: 自由度 χ が3，4，5および6のときの χ^2 分布

1.2　χ^2 検定

検定統計量が，帰無仮説の下で χ^2 分布に従うことを仮定して行う統計的検定を χ^2 検定という．いくつかの種類があるが，ここではピアソンの検定について説明する．

互いに独立である事象の観測頻度がある頻度分布に従うという帰無仮説を検定するものであり，例えばサイコロの目や通行人が男か女かなど，事象の比率が理論値に比べてどうなっているかを見るものである．ここで扱う χ^2 の値は，実は前節の式と少し異なっており，以下の式で与えられる．

$$\chi^2 = \sum \frac{(O - E)^2}{E} \tag{1}$$

ここで，O は頻度の観測値，E は帰無仮説から導かれる頻度の期待値（理論値）である．この χ^2 の値は各頻度の観測値と理論値の差を2乗し，各頻度の理論値で割って合計したものとなっている．ピアソンの検定には適合度検定と独立性検定とがあるが，まずは適合度検定について見て行こう．

1.3　適合度検定

観測された度数分布が理論分布と同じかどうかを検定するものである．例えば，標本として100人の人間がおり，これらは「男性と女性が同数いる集団からランダムに抽出された100人である」という仮説を検定するとしよう．男女の人数の観測度数と期待度数を以下のように比較する．例えば，男性が45人で女性が55人のとき，表1のようににまとめられる．

表 1: 観測結果の表

i	属性	観測度数 ν_i	期待確率 p_i	期待度数 np_i	$(\nu_i - np_i)^2/np_i$
1	male	45	1/2	50	25/50
2	female	55	1/2	50	25/50
計		$n = 100$	1	100	$\chi^2 = 1$

最後の χ^2 の値は以下の式で計算されているが，暗算でも簡単にできるであろう．

$$\chi^2 = \sum_i \frac{(\nu_i - np_i)^2}{np_i} = \frac{(45-50)^2}{50} + \frac{(55-50)^2}{50} = 1 \tag{2}$$

$n = 2$ であるので，自由度 $\phi = 1$ である．配布した χ^2 分布表を見ると，有意に異なるといえる最低ラインの α が 0.05 のときでも 3.8415 であるので，今回求められた $\chi^2 = 1$ という値は分布の右端 5%には入っていない．すなわち，帰無仮説を棄却することはできない．つまり，45 人と 55 人というこの程度の差は偶然生じることが不思議ではないということが結論付けられる．では，今年度この授業を履修している学生 83 名中に女子学生が 3 名いるというのは高校までほぼ男女同数だった集団と同質だと言えるだろうか．検定を行ってみよう．

1.4 独立性検定

２つ以上の分類基準を持つ集計表において分類基準間に関連があるかどうかを検定することを独立性検定という．例えば，以下のような例で考えてみよう．ランダムに選ばれた男女 100 人の血液型について表 2 のようなデータが得られたとする．ちなみにこのデータは下記 [1] を参考にしている．

表 2: ある集団の血液型分布

血液型	A	O	B	AB	計
male	55	22	16	7	100
female	40	32	24	4	100
計	95	54	40	11	200

このとき，帰無仮説は「調査した男女間で血液型の分布は同じである」となる．さて，統計量であるが，まずは各欄に該当する確率的な期待値を求めなければならない．期待値 E_{ij} の求め方は以下の式 (3) で求められる．

$$E_{ij} = \frac{r_i \times c_j}{n} \tag{3}$$

ここで，r_i は i 行の度数の和，c_j は j 列の和，そして n は全ての度数の和である．期待値を求めると，表 3 のようになる．

表 3: ある集団の血液型分布の理論値

血液型	A	O	B	AB	計
male	47.5	27	20	5.5	100
female	47.5	27	20	5.5	100
計	95	54	40	11	200

適合性検定で行ったのと同じように理論値からのずれの和を求め，その総和を計算する．

$$\chi^2 = \frac{(55-47.5)^2}{47.5} + \frac{(22-27)^2}{27} + \frac{(16-20)^2}{20} + \frac{(7-5.5)^2}{5.5}$$

[1] https://bellcurve.jp/statistics/course/9496.html

$$+\frac{(40-47.5)^2}{47.5}+\frac{(32-27)^2}{27}+\frac{(24-20)^2}{20}+\frac{(4-5.5)^2}{5.5}=6.639$$

今回の検定では，データの個数を見ると，2 行× 4 種類で 8 個である．自由度はそれぞれ 1 を引くので $(2-1)\times(4-1)=3$ となり，表では $\chi^2>7.8147$ で有意に分布が異なるといえることになっている．よって，この検定においても帰無仮説を棄却することはできない．

1.5　残差分析

χ^2 検定の独立性検定で危険率 $p<.05$ となると，各データ（頻度）は独立ではない，すなわち何かの依存関係があるということがわかる．そこで，どのデータに依存関係（特殊性）があるのかを見るために行うのが残差分析である．以下のような手順で行う．

まず，各欄の残差 d_{ij} を求める．この場合には，以下の式 (4) で求められる．

$$d_{ij}=\frac{f_{ij}-E_{ij}}{\sqrt{E_{ij}(1-\frac{r_i}{n})(1-\frac{c_j}{n})}} \tag{4}$$

ここで f_{ij} は i 行 j 列の頻度である．

そして，その d_{ij} は正規分布に従うことがわかっているので，正規分布の両端のすそに含まれるかどうかで有意な特殊性があるかどうかがわかる．正規分布において，$p<.05$ となるのは，絶対値が 1.96 よりも大きいときであるので，その場合にはその頻度の特殊性が有意であることがわかる．

以下の例で実際の作業を考えてみる．表 4 は，架空のアンケートであり，女性 130 人に対して，自分の体型と自分の体型に自信があるか否かの調査を行ったものとする．

表 4: 自分の体型とそれに対する自信

	自信なし	自信あり
やせ型	13	10
普通	41	17
肥満型	42	7

上記のデータに対して独立性検定を行うと $p=.02$ となり，独立では無いという結論が得られた．そこで，(4) によって求めた d_{ij} をまとめると，以下の表 5 になる．

表 5: 表 4 の残差

	自信なし	自信あり
やせ型	-2.084	2.084
普通	-0.735	0.735
肥満型	2.395	-2.395

表からは，痩せている人に自信がある人が特に多く，また，肥満型の人には自信がない人が多いという特徴的なセルを見出すことができる．なお，この表は以下のサイト [2] を参考にした．

[2] https://to-kei.net/hypothesis-testing/chi2-test-residual-analysis/