慶應義塾大学ビジネス・スクール

統計学　ノート（5）

1　カイ 2 乗分布

今まで 2 つの平均または 2 つの割合が等しいかどうかの検定の問題を，正規曲線を用いる方法で解いてきた。しかし，3 つ以上の変数がある問題をこの方法で解くことはできない。例えば，4 種類の平均を調べるとき，1 度に 2 種類ずつを比較するのは有効な解決にはならない。いくつかの割合を比較する問題を考え，いくつかの平均が等しいという仮説の検定法を与える。一般的な問題は次のように表される。

ある実験の可能な結果の数を k で表す。これらの可能な結果は k 個のマスあるいは箱で表される。実験を n 回行い，各マスに入った結果を実験全体の観測度数として表す。そのとき問題は，これらの度数が仮定したある理論から予想される度数に適合するかどうか決めることである。

このとき平均度数は期待度数とよばれ，これを e_i で表す。また観測度数は o_i で表す。

適合しているかどうか検定する一般的な方法は，観測度数と期待度数の一致の程度を測るある尺度を基にしている。この尺度はカイ 2 乗とよばれ，次の式で定義される。

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \qquad (1)$$

ここで，o_i と e_i はそれぞれ i 番目のマスの観測度数と期待度数を表し，k はマスの数を表す。

公式 (1) を見れば明らかなように，χ^2 の値は観測度数と期待度数が完全に一致したとき 0 となり，その差が大きくなれば χ^2 は大きくなる。したがって，χ^2 の値が大きくなればなるほど，実験と理論との一致はますます弱くなると考えられる。

1.1　カイ 2 乗分布

他の標本分布の場合と同様に，χ^2 の理論分布を数学的に求めることは可能である。各マスの度数がとりうる値には限界があるので，χ^2 がとりうる値にも限界がある。したがって，χ^2 の理論分布は離散形分布となるに違いない。多数の値をもつ離散型分布は，膨大な計算を必要とするので，実際的な考えとしては，2 項分布を正規分布で近似したように，離散型 χ^2 分布を連続型分布で近似させる必要がある。この近似として，カイ 2 乗分

布として知られている連続型分布が用いられる。離散型カイ 2 乗分布を近似する連続型分布もカイ 2 乗分布とよばれるが，カイ 2 乗分布という名前が用いられるのは，つねに連続型分布のほうであるから，混乱は起きないはずである。

χ^2分布の注目すべき特徴はその形はマスの個数のみで決まることである。異なるχ^2分布を区別するのに，慣例上はマスの個数kでなく自由度とよばれる母数$\nu = k - 1$が用いられる。自由度とは独立なマスの数のことである。

カイ 2 乗検定は，マスの確率が与えられなくても，“一様性”の仮定から，それが得られるような問題であれば適用することができる。

1.2 カイ 2 乗検定の制約

χ^2曲線はχ^2の真の離散型分布に対する近似でしかないため，この曲線によるχ^2検定は近似がよいときのみ用いるべきである。過去の経験と理論からどのマスも少なくとも 5 以上であれば，一般に近似は十分であることがわかっている。

もしあるマスの期待度数が 4 以下のときは，上の条件が満たされるようにそのマスを他のマスと合併するとよい。

1.3 分割表

χ^2検定のきわめて有用な応用に，2 元表における観測度数と期待度数の適合度を検定する問題がある。この 2 元表は通常，分割表とよばれている。

分割表は普通，分類基準に使う 2 つの変数間の関係を研究するためにつくられる。とくに，2 つの分類変数が全く関係がないかどうかが知りたい。χ^2検定を用いれば，2 つの分類変数が独立であるという仮説を検定できる。

この問題は前節までと異なり，観測値が各マスに落ちる確率はわかっていない。それゆえ，これまでのように各マスの期待度数を求めることができない。しかし，この困難は次のようにして克服できる。

標本をそれらが属する適当なマスに分類するという抽出実験を繰り返し行ったとする。観測された周辺合計と同じ周辺合計を与えるような繰り返し実験だけを考えるならば，そのとき各マスの期待度数を求めることができる。周辺合計は固定されているので，1 つの変数xがあるカテゴリーに入る割合は一定である。2 つの変数の間に関連がないならば，変数yのあるカテゴリーにおいても変数xがあるカテゴリーに入る割合は同じであると期待できるであろう。期待度数は変数yのあるカテゴリーに含まれる数にその割合を掛けることによって求められる。

この期待度数を用いてχ^2の値を計算する。このときのνの値はマスの数から 1 をひいたものではなく，厳密には数学的に求められる。r行c列の分割表の場合の自由度は次式で与えられる。

$$\nu = (r-1)(c-1)$$

この式は，最後の行と最後の列にあるマスの度数は他のマスの度数が与えられれば決まるという議論から導かれる。すなわち，独立な度数をもつマスの数は最後の行と最後の列とを除いたマスの数を数えれば得られる。1 つの行と 1 つの列を除いた後には，$r-1$個の行と$c-1$個の列があるので，残りのマスの数は$(r-1)(c-1)$である。

χ^2の分布は分割表のマスの数だけに依存するが，たとえ各行の割合と各列の割合が一定に保たれていても，2 つの変数が独立でないならば，標本の大きさが増すとχ^2の値も大きくなる傾向がある。したがって，nの値が大きくなるほど，独立でないことがいっそう発見されやすくなる。

参考文献

[1] 初等統計学 原書第 4 版，P. G. ホーエル，培風館 (1981)。

[2] 統計学入門，東京大学教養学部統計学教室編，東京大学出版会 (1991)。

慶應義塾大学ビジネス・スクール

Contents Works Inc.