統計学セミナー　第４回資料

北海道対がん協会　細胞診センター検査科　和田　恒之

カテゴリーデータの検定（ノンパラメトリック）

第２回で説明したデータの形式をもう一度思い出してもらいたい。統計学で扱うデータは次の三つに分類される。

1　間隔尺度
2　順序尺度
3　分類尺度

１の間隔尺度はｔ検定や分散分析で対象となるデータ形式であり、２の順序尺度はノンパラメトリック検定で扱われるデータ形式である。ただ前回の説明でも例に挙げたが体重や中性脂肪のように値そのものは間隔尺度であっても、パラメトリックの要件を満たさないときはノンパラメトリック検定として扱うことで検定が行われる。

今回は最後に残った分類尺度をどう検定するかについて説明する。ただ、間隔尺度を順序尺度として検定を行う例のように分類尺度と順序尺度もまた同様に扱うことが可能なので、両者の記述が出てくるかもしれないが了承して頂きたい。

再度、分類尺度の定義を記しておきたい。

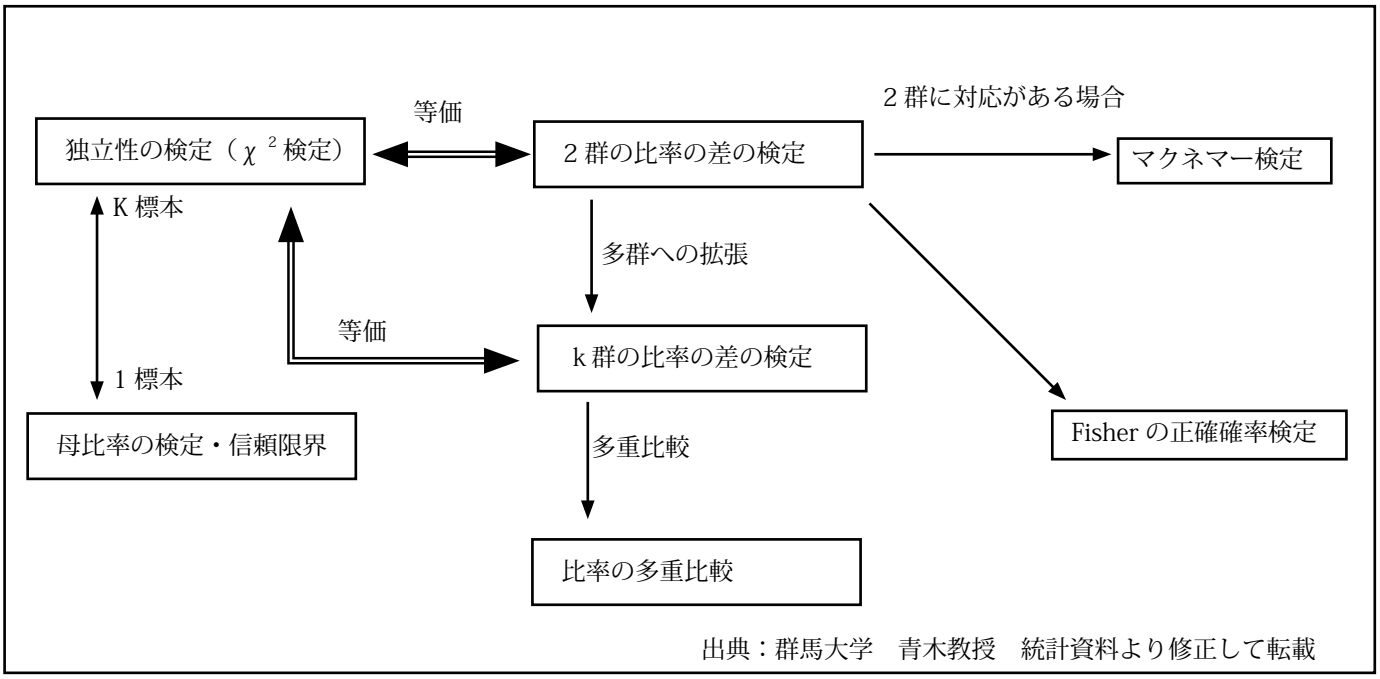
分類尺度はカテゴリー変数がその対象とされ男女の性別の違い、質問に対してのYes/Noの数など変数に順序が定義されず、問われるのはどの分類に区分けされるのかというだけのデータを指す。
間隔尺度・順序尺度がその量を計測するので計量値と呼ばれるのに対して、分類尺度ではその数を数えるので計数値と呼ばれる事もある。

検定の考え方としては、今までは帰無仮説・対立仮説にに「差がない」・「差がある」としたのに対して、帰無仮説Ｈ$_0$「観察度数（それぞれの計数値）は独立または期待度数と同じ」、対立仮説Ｈ$_1$「観察度数に偏りがある」となる。帰無仮説が支持されるのであれば､各群の観察度数は誤差以外の偏った出現をせず､どの観察度数も期待度数（理論値）どおりの出現をしている（有意な観察度数の差は無い)。対立仮説が採用されるのなら、どこかの群で誤差以外の偏った出現がある（観察度数に有意の差が見られる)。表現が異なるが意味としては今までと同様に考えて良い。

これまでの検定で触れてきたようにデータの表現として一標本、二標本、多標本、対応のある・なし、があるが、分類尺度でも同じような表現の分け方がある。第２回の資料に掲載した「データ形式と尺度による検定手法の選択例」を見て頂きたいが、標本の形に応じて分類尺度では以下のような言葉を使用する。

１標本→１要因２分類、１要因多分類
２標本→２要因２分類（２ｘ２分割）
多標本→ｌｘｍ分割（ｌは要因数、ｍは分割数）

と分けてはいるが、この関係はデータの対応があるか無いかという点から見ると簡潔に記すことができる。次ページの関係図を見ていただきたい。



この関係をよく見ると「1 標本」以外の条件で 2 群に対応がなければ χ^2 検定と 2 群の比率の差の検定、そして k 群の比率の差の検定とは互いに等価であるので χ^2 検定を覚えておけば、考え方の基礎ができる。

そこで、本資料では χ^2 検定とマクネマー検定について説明をする。

χ^2 独立性の検定（χ^2 検定） Chi-square test for independence

2 つの変数 A、B についてクロス集計表（分割表）を作成し、2 変数間に関連があるかどうかを検定する。この検定には χ^2 分布を用いるので、χ^2 検定とも呼ばれる。

例題 1

以下の表において、血液型と疾患に関連があるかを検定せよ（全く架空のデータである）

血液型	胃潰瘍	胃癌	健常者	合計
A 型	16	12	36	64
B 型	12	5	20	37
O 型	15	11	24	50
AB 型	9	2	1	12
合計	52	30	81	163

このような形の表は、日常よく見かけることが多いであろう。例題においての変数 A は血液型 , 変数 B は疾患に相当してそれぞれの変数が複数の分類に分けられている。通常、行方向を要因（k）列方向を分類（m）と呼び、k x m分割表と呼ばれる。

検定の前提である仮説は以下の通り

帰無仮説 H_0：2 変数は独立である（関連がない）

対立仮説 H_1：2 変数は独立ではない（関連がある）

各枡目の観測値、そして各枡目の期待値を計算する。R で計算を行ってみると以下の通り。

```
> chisq.test(tbl01)

        Pearson's Chi-squared test

data:  tbl5
X-squared = 13.7134, df = 6, p-value = 0.03301

Warning message:
Chi-squared approximation may be incorrect in: chisq.test(tbl5)
```

事前に tbl01 に分割表のデータを作成しておき、そのデータに対して χ^2 検定を行なった。結果はいつものように p-value を有意水準と比較すればよいので、0.033 ＜ 0.05 となるので帰無仮説は棄却されるので、「2 変数に関連がある（血液型と疾患には何らかの関連がある）」という対立仮説を採用する。

と、なるのだがもう少し独立性の検定について説明をすると、２つのカテゴリー変数が独立であるかを確かめるために行われる検定で、２つのカテゴリー変数が独立というのは、変数間に何らかの関係がないということである。（このカテゴリー変数間に関係が無いことを連関が無いという言い方をする）

＊＊ちなみに、間隔尺度のような量的変数間の関係を相関と呼び、呼び方が区別されている＊＊

例題で考えると、血液型の違い（４つの分類）と疾患の種類（健常者を含め３つの分類）、それぞれの数値の現れ方に何か関係があるのか、それとも血液型と疾患にはそんな関係は無いと言えるのかという事を検定するのだが、各セルに入っている数字は観測度数と呼ばれ実際に得られたデータになる。分割表の右側と最下列で合計となっているのは観測度数を行方向、列方向にそれぞれ合計した数字で、周辺度数と呼ばれる。集計表の右下の数値は周辺度数を合計すると得られる数値で総度数と呼ばれる。

χ^2 検定では変数の連関を調べるのに観測度数と期待度数のずれを評価する。期待度数とは帰無仮説のもとで変数間に連関がなければ、これくらいの出現数になるだろうと期待される度数のことで、期待度数は周辺度数から計算することができ以下の式から計算することができる。

分割表（クロス集計表）におけるセルの期待度数
＝（セルが属する行の周辺度数＊セルが属する列の周辺度数）/　総度数

例題の A 型・胃潰瘍患者の期待度数は（64*52)/163=20.417 となる。
全てのセルに対して（観測度数 ij －期待度数 ij)^2/ 期待度数 ij の合計をとったものを χ^2_0 とする。
*i,j は分割表のセルの位置を指定する。
これが検定統計量で自由度（k-1)x(m-1) の χ^2 分布に従うので、χ ２分布表から有意確率を求め、帰無仮説の採否を決める。
R 等、統計ソフトによる計算では p-value として有意確率が計算されるので、その値を有意水準と比較すればよろしい。

また、分割表に入る数字は必ず観察度数でなければならない、２ x ２の分割表に割合だけが示されている表からは独立性の検定は出来ない。割合しか提示されないときは母比率の推定といった別手段を使わなくてはならない。

マクネマー検定　Mcnemar's　test

　マクネマー検定は２群が対応関係にあるとき、対応のあるｔ検定の標本関係と同様だと思えばイメージがしやすいであろう。

例題１　ある政党の政策について政策発表時とその３ヶ月後、同じ有権者に調査をしてみる場合

発表時	３ヶ月後		合計
	支持	不支持	
支持	４８	２８	７６
不支持	３５	５３	８８
合計	８３	８１	１６４

例題２　２人の病理医が同じ標本を鏡検し、その評価に違いが出るかどうかを調べる場合

病理医 A	病理医 B		合計
	正常	異常	
正常	a	b	m1
異常	c	d	m2
合計	n1	n2	N

検定手順

１　説明のために記号を右図のようにする

条件 1	条件 2		合計
	特性を持つ	特性を持たない	
特性を持つ	a	b	a+b
特性を持たない	c	d	c+d
合計	a+c	b+d	n

２　帰無仮説 H0：母比率に差はない
　　対立仮説 H1：母比率に差がある
　　有意水準 α で両側検定を行う

３　標本比率の差は、(b-c)/n であり、帰無仮説の元では b=c となる。これは、ケース数＝ b+c、母比率 =1/2 の場合の二項検定（母比率の検定）と同じである。ｂ＋ｃが大きな場合は、 χ ２分布で近似ができ、この検定方法を特にマクネマーの検定と呼ぶ。

４　例題 1 を R で計算してみると

```
> mcnemar.test(matrix(c(48,28,35,53),2,2), correct=F) # 連続性の補正をしない場合
                                                    ＊ matrix（）を使いデータを直接指定している

        McNemar's Chi-squared test

data:  matrix(c(48, 28, 35, 53), 2, 2)
McNemar's chi-squared = 0.7778, df = 1, p-value = 0.3778

> mcnemar.test(matrix(c(48,28,35,53),2,2)) # 連続性の補正をする場合

        McNemar's Chi-squared test with continuity correction

data:  matrix(c(48, 28, 35, 53), 2, 2)
McNemar's chi-squared = 0.5714, df = 1, p-value = 0.4497
```

5　p-value の値が 0.377,0.449 と連続性の補正あるなしに関わらず 0.05 より大きいため帰無仮説は棄却されないので、政策発表時と三ヵ月後においての支持率に有意な差は認められないということになる。

例題 2 でも同様に a,b,c,d に当てはまる数値を入れてゆく、表中に入れる数字は以下のようになる。

a は病理医 A と B がともに正常と評価
b は病理医 A は正常、病理医 B が異常と評価
c は病理医 A は異常、病理医 B が正常と評価
d は病理医 A と b がともに異常と評価

これらの a,b,c,d の数を表に表してマクネマー検定を行なうことで、二人の病理医の評価に違いが出るかが分かる。

（注　もちろん、これは例題としての設問条件であり。仮に実在例を元に検定を行い、評価に差があったとしてもどちらが良い悪いというわけではなく。二人の評価に差が有るのか無いのかという検定の基本原則以上の結果を出しているわけではない。）

χ^2 検定や、マクネマー検定を含むカテゴリーデータを使う検定で計算の対象となっているのは、観察度数とそれをもとにする周辺度数、総度数である。あるカテゴリーに該当するかどうかの数がそれぞれの観察度数となっている以外の情報は基本的にこのデータ形式からは読み取ることはできない。検定の基本は全体に対しての実際の出現頻度（占有割合としてもいいかもしれない）が期待される出現頻度通りなのかであって、こちらのセルがあちらのセルより出現数が多い少ないということではない（多い少ないはセル内の数を見れば一目瞭然だから）。

だから、一見分割表と同じような体裁でありながら、解析の対象が観察度数では無い値であるなら、それに見合った解析方法（検定方法）を行なわなくてはならない。

例題 3

喫煙者と禁煙者（過去の喫煙者）5 人ずつに、一問につき 4 つの選択肢を持ったアンケートに答えてもらう。問題数は全部で 5 問である。設問の選択肢には集計を行い易いように 0,1,2,3 の数字を割り振り解答用紙には選択した数字を記入してもらう・・・・・

という状況で問題ごとの選択肢の選ばれ方を見るのであれば、どの選択肢に回答されるかを数え上げれば良いのでカテゴリーデータの範疇を超える事は無い (Fig.3-1)。

	選択肢 0	選択肢 1	選択肢 2	選択肢 3	合計
設問 1	2	3	1	4	10
設問 2	6	1	1	2	10
設問 3	1	5	2	2	10
設問 4	3	4	2	1	10
設問 5	0	6	1	3	10
合計	12	19	7	12	50

この表であればそれぞれのセルの観察度数を調べているので χ^2 検定を使うことに問題は無い。では選択肢の 0,1,2,3 が選択肢の分類だけでは無く、選んだ選択肢に応じて選択肢の番号と同じ点数がもらえるとしたら、0 点、1 点、2 点、3 点といった具合で最終的にこの点数の大小で何か解析したいとなったらどうだろうか？ここまで読み進められてきた方々には、既にお分かりかもしれないが、データの主従関係が設問と選択肢から、回答者と設問の選択肢に変化してしまう。(Fig.3-2)

	設問 1	設問 2	設問 3	設問 4	設問 5	合計
回答者 1	0	3	3	1	1	8
回答者 2	3	2	2	3	3	13
・・						
回答者 9	1	2	0	3	3	9
回答者 10	1	1	1	2	1	6

こうなると観察度数では無いことが分かるであろう。合計点数を扱うのであれば順序尺度、間隔尺度の検定方法を用いる必要がある。

第 4 回終了