# データ解析基礎

## 第13回 ノンパラメトリック検定

## 前回までの統計的手法の限界



- 前回までの検定の前提と手段：
  - 変数が間隔尺度である．
  - 正規分布に従っていて，分散が一定である．
  - 母集団の確率モデルを仮定し，パラメータで表現．
    → parametric method, distribution-tied method
- 以上の仮定をおくことができない場合はどうするか？
  - 名義尺度や順序尺度の場合
  - サンプル数が少ない，正規分布が仮定できない．
    → non-parametric m., distribution-free m.

## ノンパラメトリック(non-parametric) 法

3

- ノンバラメトリック法の特徴
  - 確率モデルを仮定しない．
  - パラメータ表現をしない．
  - 名義尺度や順序尺度でも使える．
- 注意すべき点
  - 考え方がパラメトリックな方法とはかなり違う．
  - パラメトリック法で扱えることがわかっている場合は，パラメトリック法の方が強力である．（対立仮説が採択されやすい）

## 適合度検定(goodness-of-fit test)

4

- 米国での昨年の日本車シェアは下左表の通りであった．一方，今年の販売実績から1000台分のサンプルを取り出したところ下右表が得られた．
  - シェアが昨年と同じであるかどうかを検定したい．

昨年度シェア

| Manu. | Share |
|---|---|
| Nissan | 18% |
| Mazda | 10% |
| Toyota | 35% |
| Honda | 37% |

今年度実績

| Manu. | Frequency |
|---|---|
| Nissan | 150 |
| Mazda | 65 |
| Toyota | 385 |
| Honda | 400 |

## 期待される頻度と実際の頻度の比較

5

- 昨年と同じ比率であれば，1000台分では下左表のような台数が期待される．

期待

| Manu. | Expected |
|---|---|
| Nissan | 180 |
| Mazda | 100 |
| Toyota | 350 |
| Honda | 370 |

実績

| Manu. | Observed |
|---|---|
| Nissan | 150 |
| Mazda | 65 |
| Toyota | 385 |
| Honda | 400 |

- 期待される台数（左表）と実績（右表）を比較すれば，二つの年が同じであるかどうかを判断できそうである．

## 適合度検定での検定統計量

6

- 「期待と実際に違いがない」という帰無仮説の下で，以下の量が自由度 $K$-1 のカイ2乗分布に従う．

$$\chi^2 (K\text{-}1) \quad = \quad \Sigma_i \, (O_i - E_i)^2 / E_i$$

$K$: グループ数

$O_i$: 観測された頻度，$E_i$：予測される頻度

- あとの手続きはこれまでと同じ．
- この量は両者の違いが大きいほど大きな値になるので，常に片側検定を行なう．

## 例題の答え

7

- 棄却域：

  有意水準を5%とすると，自由度3 (= 4 - 1)のカイ2乗分布の数表から，棄却域は$\chi^2 > 7.815$である．

- 検定統計量

  前スライドの式に実際に値を代入して計算すると，$\chi^2 = 23.18$となる．よって，帰無仮説は棄却される．

- p値

  $p = 0.00003704$であるから，この検定は有意水準を0.1%に設定しても有意になる．

## 適合度検定の性質

8

- 特徴
  - 期待度数と実際の度数を比較するだけなので，分布を仮定する必要がない．
- 注意すべき点
  - グループは互いに排他的で，かつすべての場合を網羅している必要がある．
  - 実験（試行）は独立でなければならない．
  - 期待度数の和と実際度数の和が等しい必要がある．
  - 期待度数は少なくとも5程度ないとうまくいかない．
  - 自由度が1の場合はYatesの補正を加える．

## 独立性の検定

9

- 男女各100人のサンプルを対し，次の結果が得られた．

| | 自民党 | 民主党 | それ以外 |
|---|---|---|---|
| 男性 | 40 | 10 | 50 |
| 女性 | 60 | 20 | 20 |

- 男性・女性間で支持政党の割合に差があるかどうかを知りたい．

  つまり，性別と支持政党が独立であるか（無関係であるか）どうかを知りたい．

## 適合度検定との関係

10

- 男女あわせた全体度数は，性別に依らない全体での度数であるから，「期待度数」とみなせる．
- 男女それぞれの度数は「観測度数」とみなせる．
  - 適合度検定と同じ考え方で検定ができそうである．
- 両者に差がない（独立である）という仮定の下で，次の量が自由度 (グループ数 - 1)×(候補数 - 1) のカイ2乗分布に従うと考えて検定を行なう．

$$\chi^2 = \Sigma_i (O_i - E_i)^2 / E_i$$

## 例題の答え：

11

- 期待度数と実際の度数を表にすれば，次のようになる．

| | 自民党 | 民主党 | それ以外 |
|---|---|---|---|
| 男性 | 40 | 10 | 50 |
| 女性 | 60 | 20 | 20 |
| 期待度数 | 50 | 15 | 35 |

- 自由度2 ( = (3-1) x (2-1))のカイ2乗分布の数表から，有意水準1%の場合の棄却域は $\chi^2 > 9.210$である．
- 検定統計量を計算すると20.19となる．
  - したがって，性別と支持政党は独立ではない．

## データが順序尺度である場合の検定

12

- 符号検定：
  - 互いに対応している二つの集団を比較する．
- Mann-Whitney のU検定
  - 無関係な二つの集団を比較する．
- Kruskal-Wallis の検定
  - 無関係な三つ以上の集団を比較する．
    （one-way ANOVAのnon-parametric版）
- いずれも「比較する母集団の分布が同一であること」を帰無仮説として検定を行なう．

## 符号検定

13

- 対応する二つの集団のデータを直接比較してどちらが大きいかを調べ，それに応じて正負の符号を付ける．
- 二つの集団に違いがなければ，符号が正になるか負になるかは $p = 0.5$ のBernoulli 試行とみなせる．
- したがって，$N$個のデータ対の内符合が正の対の数は二項分布に従うはずである．
- 帰無仮説の下で二項分布に従うとして棄却域を設定．
- データが多いときは，中心極限定理により正規分布．
  - 復習：平均は $N / 2$，標準偏差は $\mathrm{sqrt}(N)/ 2$ ．

## 例題：

14

- 新しい教育プログラムに効果があるかどうかを調べるため，24人の学生を二つに分けて，新プログラムと旧プログラムの下で指導を受けさせた．
- 成績がほぼ同程度の学生を対にして，指導後の成績を比較したところ，次のようになった．新プログラムの効果（良い場合も悪い場合も含めて）があるかどうかを符号検定で検定せよ．

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 新: | 5, | 8, | 5, | 4, | 8, | 9, | 7, | 6, | 6, | 8, | 10, | 8 |
| 旧: | 5, | 10, | 7, | 7, | 8, | 10, | 7, | 9, | 5, | 9, | 9, | 10 |

## 例題の解答：

15

- 対応する学生の成績を比較する．
- 差がない対が三つあるので，これを除外する．
- 新プログラムの方が成績がよい対の数は 2 である．
  - 「旧プログラムがよい対は7である」としてもよい．
- $N$=9, $p$=0.5 の二項分布の累積分布より，有意水準5%の両側検定の棄却域は，$m$ = 0, 1, 8, 9となる．
- 2は棄却域の外なので，帰無仮説は棄却されない．
  - 「新プログラムの効果はない」と結論される．

## Mann-Whitney のU検定

16

- 二つの集団の差が，一つの集団を二つに分けたときに生じる差に比べて統計的に大きいかどうかを調べる．
- 統計量として，以下のように定義されるUを使う．
  - グループ1の方が全体として値が大きいとする．
  - グループ1の個々のデータに着目し，グループ2のデータの中でそれよりも小さなものの数を数える．
  - 上で求めた数をグループ1のデータすべてについて足しあわせたものがUである．

## U統計量

17

- 二つのグループのデータ数をそれぞれ$N_1$, $N_2$とすると，Uの値は 0 から $N_1N_2$ までの値をとる．
- 二つのグループのデータが同じように分布していれば，Uは $N_1N_2 / 2$ に近い値をとると思われる．
- Uが $N_1N_2 / 2$ に比べて大きいかどうかを調べればよい．具体的には，Uの数表を用いて棄却域を決定する．

## U統計量の系統的な求め方

18

- Uを系統的に求める方法として，以下の方法がある．
- 二つのグループをあわせて小さな順に並べたときの順位をグループ1 について合計したものを$R_1$とする．
- 同様にして，順位をグループ2 について合計したものを$R_2$とする．
- $R_1$と$R_2$を使うと，Uは次の式で与えられる．

$$U = R_1 - N_1(N_1 + 1) / 2$$
$$= N_1N_2 + N_2(N_2 + 1) / 2 - R_2$$

## 例題：

19

- 次の二群の分布に違いがあるかどうかを検定せよ．

| | | | | | |
|---|---|---|---|---|---|
| グループ1: | 14 | 17 | 18 | 25 | |
| グループ2: | 7 | 13 | 15 | 16 | 12 |

- 方法１：先ほどのアルゴリズムを使ってみる．
  - 14より小さなグループ2のデータ数は3．
  - 17より小さなものは5個．
  - 18より小さなものは5個．
  - 25より小さなものは5個．

  よって，U= 3 + 5 + 5 + 5 = 18.

## 例題（続き）

20

- 方法2：すべてのデータを並べて順序をつける．

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| 7 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 25 |

  - $R_1$ = 4 + 7 + 8 + 9 = 28 より

    $U = R_1 - N_1(N_1 + 1)/2$ = 28 - 4 x (4+1) / 2 = 18

  - または，$R_2$ = 1 + 2 + 3 + 5 + 6 = 17であるから，

    $U = N_1N_2 + N_2(N_2 + 1)/2 - R_2$ = 18

  - $N_1$ = 4, $N_2$ = 5の数表から有意水準5%の両側検定の棄却域を求めると，0, 1, 19, 20 となる．
- したがって，帰無仮説は棄却されない．

## 補足：

21

- データ数が多いときUは正規分布に従うので，正規分布の数表を使って検定できる．
  - 平均は $N_1N_2 / 2$
  - 標準偏差は $\mathrm{sqrt}( N_1 N_2 (N_1 + N_2 + 1) / 12 )$

  となる．

## Kruskal-Wallisの検定

22

- 母集団が三つ以上の場合の方法である．
- 以下の量Hが自由度 $K - 1$ のカイ2乗分布に従う．

$$H = \frac{12}{N(N+1)} \Sigma_j \frac{R_j^2}{N_j} - 3(N+1)$$

$K$： グループの数

$N_j$ : グループ $j$ に含まれるデータの数

$N$：全データ数

$R_j$：グループ $j$ に含まれるデータの順位の和

## 例題：

23

- 株式投資に対する考え方が20代，40代，60代でどのように違うかを調べたい．
- 25人の人に株式投資に関するアンケートを行なって得点を集計したところ，以下のデータが得られた．
- 世代間での考え方に違いがあるかどうかを検定せよ．

| | |
|---|---|
| 20代 | 48, 42, 40, 46, 35, 39, 32, 41 |
| 40代 | 28, 33, 26, 34, 29, 36, 31, 22, 21, 17 |
| 60代 | 15, 19, 20, 25, 18, 27, 16 |

## 解答：

24

- 準備
  - $N_1 = 8$, $N_2 = 10$, $N_3 = 7$,  $N = 25$, $K = 3$
- データに順位をつけて，順位和を求める．．
  - $R_1 = 15 + 18 + 20 + 21 + 22 + 23 + 24 + 25 = 168$
  - $R_2 = 3 + 7 + 8 + 10 + 12 + 13 + 14 + 16 + 17 + 19 = 119$
  - $R_3 = 1 + 2 + 4 + 5 + 6 + 9 + 11 = 38$

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 25 | 26 |
| 10 | 27 | 28 | 29 | 31 | 32 | 33 | 34 | 35 | 36 | 39 |
| 20 | 40 | 41 | 42 | 46 | 48 | | | | | |

## 解答（続き）



- Hを求める．

$$H = \frac{12}{N(N+1)} \sum_j \frac{R_j^2}{N_j} - 3(N+1)$$

$$= \frac{12}{25 \times 26} \left[ \frac{168^2}{8} + \frac{119^2}{10} + \frac{38^2}{7} \right] - 3 \times 26$$

$$= 14.71$$

- 自由度2のカイ2乗分布の数表から，有意水準5%の棄却域は $H > 5.991$．
- したがって，「三つのグループの間には差がある」という結論が得られる．