統計学 授業資料

# 第1回　CochranのQ検定

正確法と、それ以外の方法

児玉 靖司
法政大学

2026年11月22日

**この回の目標**

1. 同じ人（同じ対象）を三つ以上の条件で測った「成功／失敗」のデータを、CochranのQ検定で比べられる。
2. p値を*正確法*（すべての場合を数え上げる方法）で求める考え方が分かる。
3. 正確法以外の方法（$\chi^2$近似、モンテカルロ法）を知り、どれを使えばよいかを判断できる。
4. McNemar検定・Friedman検定など、似た場面で使う方法との関係が分かる。

## 目次

**1　どんな問題か** 2

**2　統計量Qの計算** 3

**3　正確法：すべての場合を数え上げる** 3

3.1　p値とは . . . 3
3.2　帰無仮説の下で、何が「偶然」なのか . . . 4
3.3　場合の数を数える . . . 4
3.4　Qの分布 . . . 4
3.5　p値 . . . 4

**4　正確法以外の方法** 5

4.1　$\chi^2$近似 . . . 5
4.2　モンテカルロ法（無作為な並べ替え） . . . 6
4.3　方法の比較 . . . 6

**5　似た場面で使う方法** 6

5.1　McNemar検定（条件が2つのとき） . . . 6
5.2　事後比較（どの組に差があるか） . . . 7

5.3 Friedman検定（結果が順位や点数のとき） . . . . . 7
5.4 Wilcoxonの符号付き順位検定（2条件で、点数のとき） . . . . . 7
5.5 対応がないとき . . . . . 7
5.6 モデルを使う方法 . . . . . 8
5.7 まとめの表 . . . . . 8

**6 まとめ** **8**

**7 練習問題** **8**

# 1 どんな問題か

**例題。** 10人の学生が、同じ単元の3つの問題A・B・Cを解いた。正解を1、不正解を0とすると、結果は表1のようになった。*3つの問題の難しさに差はあるか。*

表 1: 例題のデータ。$R_i$は学生ごとの正解の数、$C_j$は問題ごとの正解の数。

| 学生 | A | B | C | $R_i$ |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 3 |
| 2 | 1 | 1 | 0 | 2 |
| 3 | 1 | 0 | 0 | 1 |
| 4 | 1 | 1 | 0 | 2 |
| 5 | 1 | 1 | 1 | 3 |
| 6 | 1 | 0 | 0 | 1 |
| 7 | 0 | 0 | 0 | 0 |
| 8 | 1 | 1 | 0 | 2 |
| 9 | 1 | 0 | 0 | 1 |
| 10 | 1 | 0 | 1 | 2 |
| $C_j$ | 9 | 5 | 3 | $N = 17$ |

正解率はAが90%、Bが50%、Cが30%で、差があるように見える。しかし、学生は10人しかいない。この差が偶然でも起こる程度かどうかを調べたい。

**データの特徴。** このデータには、次の二つの特徴がある。

- **結果が2値**（正解か不正解か）である。
- **対応がある。**同じ学生が3問すべてを解いている。よくできる学生（学生1・5）もいれば、どれも解けない学生（学生7）もいる。学生の力の差は大きいので、問題の差は*学生の中*で比べたい。

2値で対応のある三つ以上の条件を比べる代表的な方法が、**CochranのQ検定**である（Cochran, 1950）。

**仮説。**

- 帰無仮説$H_0$：3問の難しさは同じである（どの問題も、正解になる確率は等しい）。
- 対立仮説$H_1$：少なくとも一つの問題の難しさが違う。

## 2　統計量Qの計算

条件（問題）の数を$k$、対象（学生）の数を$n$とする。$x_{ij}$を学生$i$の問題$j$の結果（1か0）として、

$$C_j = \sum_i x_{ij} \quad \text{（問題ごとの正解数）}, \qquad R_i = \sum_j x_{ij} \quad \text{（学生ごとの正解数）},$$

$$N = \sum_j C_j = \sum_i R_i.$$

CochranのQ統計量は

$$\boxed{Q = \frac{(k-1)\left(k\sum_j C_j^2 - N^2\right)}{kN - \sum_i R_i^2}}$$

である。

**意味。**　分子の$k\sum C_j^2 - N^2$は、問題ごとの正解数$C_j$のばらつき（の$k^2$倍の分散）である。難しさが同じなら$C_j$はそろい、分子は0に近くなる。分母は、学生の中で正解と不正解が分かれる程度で、尺度をそろえる役目をする。

**例題での計算。**　$k = 3$、$N = 17$である。

$$\sum_j C_j^2 = 9^2 + 5^2 + 3^2 = 115, \qquad \sum_i R_i^2 = 9+4+1+4+9+1+0+4+1+4 = 37,$$

$$Q = \frac{2\,(3 \times 115 - 17^2)}{3 \times 17 - 37} = \frac{2 \times 56}{14} = 8.0.$$

$Q$が大きいほど、問題の間の差が大きい。では、$Q = 8.0$は「偶然では起こりにくい」ほど大きいのだろうか。それを決めるのがp値である。

**全部正解・全部不正解の学生は、Qに影響しない**

学生1・5（全部正解）と学生7（全部不正解）を取り除いて計算し直してみよう。$C = (7, 3, 1)$、$N = 11$、$\sum R_i^2 = 37 - 9 - 9 - 0 = 19$なので、

$$Q = \frac{2\,(3 \times 59 - 11^2)}{3 \times 11 - 19} = \frac{2 \times 56}{14} = 8.0$$

と、同じ値になる。全部正解の行を1行加えると、分子の$k\sum C_j^2$と$N^2$は同じだけ増え、分母の$kN$と$\sum R_i^2$も同じだけ増えるので、差は変わらない（全部不正解の行は何も変えない）。全部正解・全部不正解の行は、*どの問題が難しいかについて情報を持たない*のである。

## 3　正確法：すべての場合を数え上げる

### 3.1　p値とは

p値は、「帰無仮説が正しいとしたとき、観測した結果と同じか、それより極端な結果が出る確率」である。小さいほど、帰無仮説と観測が合わない。よく使われる基準は5%（0.05）である。

### 3.2　帰無仮説の下で、何が「偶然」なのか

3問の難しさが同じなら、ある学生の「正解2つ・不正解1つ」が*どの問題*で起きたかは偶然である。例えば学生2の$(1,1,0)$は、同じ確率で$(1,0,1)$や$(0,1,1)$でもあり得た。つまり、

> *各学生の正解の数$R_i$はそのままにして、どの問題に正解を割り当てるかを入れ替えても、帰無仮説の下では同じ確率で起こる。*

これが正確法（**並べ替え検定**、permutation test）の考え方である。

### 3.3　場合の数を数える

各学生について、入れ替え方の数は次のとおりである。

- 全部正解$(1,1,1)$、全部不正解$(0,0,0)$：入れ替えても同じなので1通り（学生1・5・7）。
- 正解1つ：どの問題が正解かで3通り（学生3・6・9）。
- 正解2つ：どの問題が不正解かで3通り（学生2・4・8・10）。

成否が分かれた学生は7人なので、全体の並べ替えは$3^7 = 2187$通りである。帰無仮説の下では、これらはすべて同じ確率$1/2187$で起こる。

### 3.4　Qの分布

入れ替えても$R_i$は変わらないので、$Q$の分母$kN - \sum R_i^2 = 14$も変わらない。変わるのは$\sum C_j^2$だけである。例題では$Q = (3\sum C_j^2 - 289)/7$となり、$Q$は$1/7$刻みの値しかとらない。2187通りすべてについて$Q$を計算すると、表2と図1のようになる。

表 2: 帰無仮説の下でのQの正確な分布（2187通り）。右の2列は「その値以上」の確率。

| $Q$ | 通り | 確率 | $P(Q \geq q)$ 正確法 | $\chi^2_2$近似 $e^{-q/2}$ |
|---:|---:|---:|---:|---:|
| $2/7 \approx 0.29$ | 639 | 0.292 | 1.000 | 0.867 |
| $8/7 \approx 1.14$ | 432 | 0.198 | 0.708 | 0.565 |
| $14/7 = 2.00$ | 594 | 0.272 | 0.510 | 0.368 |
| $26/7 \approx 3.71$ | 252 | 0.115 | 0.239 | 0.156 |
| $32/7 \approx 4.57$ | 90 | 0.041 | 0.123 | 0.102 |
| $38/7 \approx 5.43$ | 114 | 0.052 | 0.082 | 0.066 |
| $50/7 \approx 7.14$ | 18 | 0.008 | 0.030 | 0.028 |
| $56/7 = 8.00$ | 24 | 0.011 | 0.022 | 0.018 |
| $62/7 \approx 8.86$ | 18 | 0.008 | 0.011 | 0.012 |
| $74/7 \approx 10.57$ | 6 | 0.003 | 0.003 | 0.005 |
| 計 | 2187 | 1.000 | | |

### 3.5　p値

観測した$Q = 8.0$以上になるのは、$24 + 18 + 6 = 48$通りである。したがって

$$p = \frac{48}{2187} = 0.022.$$

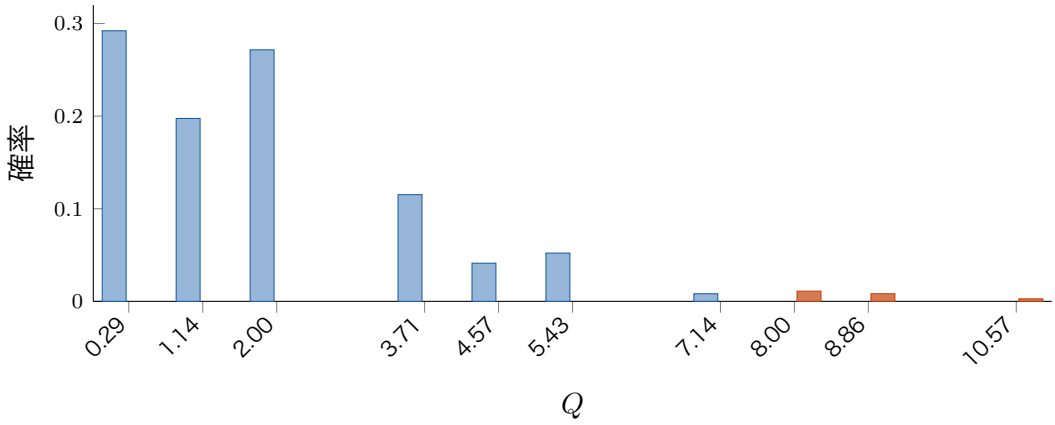



図 1: 帰無仮説の下でのQの正確な分布。オレンジは観測値$Q = 8.0$以上の部分で、その確率の和がp値である。

5%水準で有意である。「3問の難しさは同じ」という仮説は、データと合わない。

**正確法の手順のまとめ**

1. 観測データから$Q$を計算する。
2. 各対象（行）の中で、成否の並びを入れ替えたすべての組合せを作る。
3. それぞれで$Q$を計算し、観測値以上になる割合をp値とする。

**正確法の長所と短所。**

- **長所**：近似を使わないので、対象が少なくても正しいp値が得られる。
- **短所**：組合せの数は、成否の分かれた行が$m$個なら最大で$\binom{k}{1}^m$程度に増える。$k = 3$で$m = 20$なら約35億通りで、そのままでは数え切れない。工夫（列和の分布を行ごとに畳み込む動的計画法）で速くできるが、大きなデータでは次節の方法を使うことが多い。
- **p値がとびとびになる。**$Q$のとる値が限られるので、p値も限られた値しかとらない。そのため、5%ちょうどの検定を作れず、やや*保守的*（有意になりにくい）になる。この点を和らげる**mid-p値**（観測値と同じ値の確率を半分だけ数える）もある。例題では$(48 - 24/2)/2187 = 0.016$である。

# 4　正確法以外の方法

## 4.1　$\chi^2$近似

対象の数が十分に多いとき、帰無仮説の下で$Q$は**自由度**$k-1$**の**$\chi^2$**分布**に近づく。そこで、

$$p \approx P(\chi^2_{k-1} \geq Q)$$

とする。自由度2の場合は$P(\chi^2_2 \geq q) = e^{-q/2}$と簡単に書けるので、例題では

$$p \approx e^{-8.0/2} = e^{-4} = 0.018$$

である。自由度が3以上なら、$\chi^2$分布表か、表計算ソフトの関数で求める。Excelなら`=CHISQ.DIST.RT(Q,k-1)`である。

**近似の良し悪し。**　表2の右の2列を比べると、正確な確率と近似は、$Q$が小さいところでは大きくずれている（例えば$Q \geq 2$で0.510対0.368）。観測値$Q = 8.0$の付近では0.022対0.018と近いが、近似の方が小さい。近似は、少ないデータでは有意になりやすい方向にずれることがある。

> **目安：**成否の分かれた行（全部1でも全部0でもない行）が少ないときは、$\chi^2$近似を使わず、正確法か次のモンテカルロ法を使う。統計ソフトの多くは既定で$\chi^2$近似を出すので、出力をそのまま報告する前に、分かれた行の数を確かめること。

### 4.2　モンテカルロ法（無作為な並べ替え）

正確法の組合せが多すぎるときは、すべてを数える代わりに、*無作為に選んだ*並べ替えで代用する。

1. 各行の中で、成否の並びを無作為に入れ替えたデータを作る。
2. その$Q$を計算する。
3. これを$B$回（例えば1万回）くり返し、観測値以上になった回数を$b$とする。
4. $p \approx b/B$（または、観測データ自身も一つの並べ替えとして数える$(b+1)/(B+1)$）。

例題で$B = 10000$回くり返すと、$b = 203$で$p \approx 0.020$だった。$p$の推定の標準誤差は$\sqrt{p(1-p)/B} \approx 0.0014$なので、正確な値0.022とよく合っている。$B$を増やせば、いくらでも正確法に近づけられる。

### 4.3　方法の比較

表 3: 例題での3つの方法のp値。

| 方法 | p値 | 計算の手間 | 使う場面 |
|---|---|---|---|
| 正確法（全数え上げ） | 0.022（48/2187） | 組合せが多いと重い | 対象が少ないとき。答えが一つに決まる |
| モンテカルロ法 | 0.020（$B$=10000） | $B$回の計算 | 組合せが多すぎて数え切れないとき |
| $\chi^2$近似 | 0.018 | 一瞬 | 成否の分かれた行が十分に多いとき |

## 5　似た場面で使う方法

### 5.1　McNemar検定（条件が2つのとき）

条件が2つ（$k = 2$）なら、CochranのQは**McNemar検定**と同じになる。2つの問題AとCを比べると、成否が分かれた学生は、

- Aだけ正解：$b = 6$人（学生2・3・4・6・8・9）、
- Cだけ正解：$c = 0$人

である（学生10は両方正解）。McNemarの統計量は$(b-c)^2/(b+c) = 36/6 = 6.0$で、これは2列だけで計算したQと一致する。

**正確法**では、分かれた$m = b + c = 6$人それぞれで、どちらが正解するかが半々（二項分布Bin$(6, 1/2)$）と考えて、

$$p = 2 \times P(b' \leq 0) = 2 \times \frac{1}{2^6} = 0.031.$$

$\chi^2$**近似**（自由度1）では$p \approx 0.014$で、ここでも近似の方が小さい。

## 5.2 事後比較（どの組に差があるか）

Qが有意になっても、分かるのは「どこかに差がある」ことだけである。どの組に差があるかは、組ごとにMcNemar検定（正確法）を行い、**多重比較の補正**をする。例題の3組では、

表 4: 組ごとのMcNemar検定（正確法）とHolmの補正。

| 組 | 前だけ正解 | 後だけ正解 | p値 | Holmの補正 | 判定（5%） |
|---|---|---|---|---|---|
| A対C | 6 | 0 | 0.031 | $0.031 \times 3 = 0.094$ | 有意でない |
| A対B | 4 | 0 | 0.125 | $0.125 \times 2 = 0.250$ | 有意でない |
| B対C | 3 | 1 | 0.625 | $0.625 \times 1 = 0.625$ | 有意でない |

となる（Holmの方法：p値を小さい順に並べ、$k'$組なら順に$k'$倍、$k'-1$倍、…して比べる）。全体のQは有意だったのに、組ごとにはどれも有意にならない。対象が10人では、*一つの組の差を言い切るには足りない*ということである。これは珍しいことではない。

## 5.3 Friedman検定（結果が順位や点数のとき）

結果が2値でなく、点数や順位なら、**Friedman検定**を使う。各行の中で値に順位をつけ、列ごとの順位の和を比べる方法である。

実は、*2値のデータに、同順位の補正をしたFriedman検定を当てはめると、CochranのQと一致する*。例題で、各学生の中で順位をつけ（同じ値には平均の順位）、補正つきの式で計算すると8.0になる。ただし、同順位の補正を*し*ない式を使うと4.2になり、大きく違う。2値のデータにFriedman検定を使うときは、補正があるかどうかに注意する。

## 5.4 Wilcoxonの符号付き順位検定（2条件で、点数のとき）

条件が2つで、結果が点数なら、**Wilcoxonの符号付き順位検定**を使う。McNemar検定との関係は、Friedman検定とCochranのQの関係と同じである。

## 5.5 対応がないとき

3つの問題を*別々の学生*が解いた場合は、対応がない。このときは、$3 \times 2$の分割表（問題×正解・不正解）に対する$\chi^2$**検定**か、その正確法である**Fisherの正確検定**を使う。対応があるのに対応のない方法を使うと、学生の力の差がそのまま誤差に入り、差を見落としやすくなる。

### 5.6　モデルを使う方法

学生の属性（学年など）を一緒に考えたいとき、または欠けた値があるときは、**一般化線形混合モデル**（学生を変量効果とするロジスティック回帰）や**一般化推定方程式**（GEE）を使う。CochranのQより柔軟だが、対象が少ないと推定が不安定になる。

### 5.7　まとめの表

表 5: 結果の種類と、対応の有無による方法の選び方。

| | 対応あり・2条件 | 対応あり・3条件以上 | 対応なし |
|---|---|---|---|
| 2値（成功・失敗） | McNemar検定 | **CochranのQ検定** | $\chi^2$検定、Fisherの正確検定 |
| 順位・点数 | Wilcoxonの符号付き順位検定 | Friedman検定 | Mann–Whitney、Kruskal–Wallis |

いずれの方法にも、正確法（並べ替え）・モンテカルロ法・近似法の三つの求め方がある。考え方はこの回と同じである。

## 6　まとめ

- CochranのQ検定は、*2値・対応あり・3条件以上*のデータで、条件の間に差があるかを調べる。
- 正確法は、各行の中で成否を入れ替えたすべての場合を数え、観測値以上の割合をp値とする。例題では$48/2187 = 0.022$。
- $\chi^2$近似（0.018）は手軽だが、対象が少ないとずれる。組合せが多すぎるときはモンテカルロ法（0.020）を使う。
- 条件が2つならMcNemar検定、点数ならFriedman検定。Qが有意でも、どの組に差があるかは多重比較の補正をして別に調べる。

## 7　練習問題

**問1.**　7つの事例で、方式X・Y・Zの成功（1）と失敗（0）が次のようだった。

| 事例 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| X | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| Y | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| Z | 1 | 1 | 1 | 1 | 1 | 1 | 0 |

(1) $C_j$、$R_i$、$N$を求め、$Q$を計算せよ。

(2) 正確法で、並べ替えは何通りあるか。

(3) $\chi^2$近似のp値を求めよ。

(4) 正確法のp値は$6/81$である。(3)と比べて何が言えるか。

**問2.** 問1のYとZについて、McNemar検定（正確法）のp値を求めよ。

**問3.** 次の場面では、どの検定を使うか。

(1) 30人に、同じ文章を3種類のフォントで読ませ、読みやすさを5段階で答えさせた。

(2) 3つのクラス（別々の学生）で、同じ問題の正解・不正解を比べた。

(3) 12台のロボットで、2つのプログラムのそれぞれについて、作業の成功・失敗を記録した。

### 解答

**問1.** (1) $C = (2, 4, 6)$、$R = (3, 2, 1, 2, 3, 1, 0)$、$N = 12$、$\sum C_j^2 = 56$、$\sum R_i^2 = 28$。$Q = 2(3 \times 56 - 144)/(36 - 28) = 48/8 = 6.0$。 (2) 成否の分かれた事例は2・3・4・6の4つで、$3^4 = 81$通り。 (3) $p \approx e^{-6.0/2} = e^{-3} = 0.050$。 (4) 正確法の$6/81 = 0.074$より小さい。近似では5%水準ぎりぎりに見えるが、正確には有意でない。事例が少ないと、近似はp値を小さく見積もることがある。

**問2.** Yだけ成功が0件、Zだけ成功が2件（事例3・6）。$m = 2$で$p = 2 \times (1/2)^2 = 0.50$。分かれた事例が2件しかないので、2件ともZが勝っても有意にはなりえない。

**問3.** (1) Friedman検定（対応あり・3条件・順序）。(2) $3 \times 2$分割表の$\chi^2$検定、またはFisherの正確検定（対応なし）。(3) McNemar検定（対応あり・2条件・2値）。

## 付録：計算のプログラム（Python）

例題の数値は、次のプログラムで確かめた。xを問1のデータに変えれば、解答の値が得られる。

```
import itertools, math, random

X = [[1,1,1],[1,1,0],[1,0,0],[1,1,0],[1,1,1],
     [1,0,0],[0,0,0],[1,1,0],[1,0,0],[1,0,1]]   # 表1

def Q(X):
    k = len(X[0]); C = [sum(r[j] for r in X) for j in range(k)]
    R = [sum(r) for r in X]; N = sum(R)
    return (k-1) * (k*sum(c*c for c in C) - N*N) / (k*N - sum(v*v for v in R))

q = Q(X)                                          # 8.0
print('chi2 approx (k=3):', math.exp(-q/2))       # 0.0183

rows = [sorted(set(itertools.permutations(r))) for r in X]   # 行ごとの並べ替え
all_ = list(itertools.product(*rows))                          # 2187 通り
ge = sum(Q(list(c)) >= q - 1e-9 for c in all_)
print('exact:', ge, '/', len(all_), '=', ge/len(all_))        # 48/2187 = 0.0219

random.seed(1); B = 10000                                      # モンテカルロ法
b = sum(Q([list(random.choice(r)) for r in rows]) >= q - 1e-9 for _ in range(B))
print('Monte Carlo:', b/B)                                     # 0.0203
```

**参考文献**


- W. G. Cochran, “The comparison of percentages in matched samples,” *Biometrika*, 37(3/4), 256–266, 1950.
- Q. McNemar, “Note on the sampling error of the difference between correlated proportions or percentages,” *Psychometrika*, 12(2), 153–157, 1947.
- M. Friedman, “The use of ranks to avoid the assumption of normality implicit in the analysis of variance,” *Journal of the American Statistical Association*, 32(200), 675–701, 1937.