パーミュテーション検定について

松　井　　敬*

1．はじめに

R.A.Fisher による「婦人と紅茶の飲み分け」の話は彼の著書［3］に見られるよく知られた挿話[*1]だが，実験の場における無作為化の重要性を述べつつ統計的有意性の判断の方法を示す好例である[*2]．これを拡張した形の統計的方法が 2×2 分割表による Fisher 精密検定（exact test）である．これは，分割表においてサンプルサイズが比較的小さい場合に適用され，最小セルの値を変化させて確率を計算しそれらの値をもとに有意性を検討している．

多くの場合 2×2 分割表ではサンプルサイズが小さいとはいえず，この場合精度の高い計算はかなりやっかいなことになる．組合せ（階乗）の計算が多くなるからである．実際筆者が学生の頃でさえ，よく知られた「バーローの数表」に，三角関数や対数の値のみならず平方根の値などが数表化され利用に供されていたほどで，計算尺も有効な計算手段として用いられていた．電卓ひとつでこれらの値を算出できる今の時代とは計算環境がまるで異なっていたわけである．このため大標本における分割表では χ^2 による近似法が考えられ，もっぱらその方法が用いられてきた．その一方で，精密検定の方法はむしろ例外的な方法としてとり扱われてきていた．

Fisher は上述の著書の中で連続型のデータの例として C. ダーウィンのデータに対し，平均の差にかかわるパーミュテーション検定を行っている．このため，32,768 通りの並べかえによって判定基準となる permutation distribution（P-分布，後述）を得ている．これが連続データにパーミュテーション検定を適用した嚆矢とされているが，これらはすべて手計算でなされていて大変な計算量といえる．Ludbrook & Dudley [7] は， Fisher がこのデータ解析の後パーミュテーション検定について言及しなくなったのは，この計算の煩雑さにうんざりしたからであろうと述べている．

Fisher による上記のようなパーミュテーション検定に対する言及以降，パーミュテーション検定が再び強く意識されてきたのは計算環境が劇的に変化してきたことを考慮して， 1900 年代後半の頃からであり，実際にはパソコンなどが導入された 1990 年代に入ってからである．

統計的仮説検定は多くの場合想定した母集団分布の母数に対し，まずそれを検定するための統計量を考える．次に，その統計量の分布を漸近的な場合も含め理論的に捉え，その上で母数に対して与えられた仮説を実際のデータをもとに検討する（population model）．あるいは実験対象に無作為に処理を割り当て，処理群と非処理群との間の差を検出しようとする（randomization model）．

このとき，あらゆる可能な場合について統計量がとりうる値を求め，その分布が対応する確率を計算し，当該の仮説についての判断を行おうとするのがパーミュテーション検定（permutation test）である．これは「並べかえ検定」と訳されることがある

* 獨協大学名誉教授

*1 ある婦人が紅茶の入れ方について (a) 「紅茶を先に入れて後からミルクを注いだ」か (b) 「ミルクを先にし後から紅茶を入れた」かを飲み分けられると主張した．フィッシャーは (a) と (b) の紅茶を 4 カップずつ 8 カップ用意し、ランダムに配置し飲み分けてもらうことにした．Moore & Condit [9] ではこれをパーミュテーション検定の立場から論じている．

*2 これについては Salsburg [10] によるこのタイトルの著作もあり，興味深い話が展開されている

が，名称に少し違和感もあるので，本論文では以下 P-検定 と記すことにする．

統計学の推測理論がこれまで多くの年月を重ねる中で，データを扱うに際しての環境は大きく変わってきている．ビッグデータとよばれる環境もあるし，かつて「少数例のまとめ方」として喧伝されたように，特に生物医薬分野ではデータの数は必ずしも大きいとは言えない—実はデータの数が少ない方が分析法の選択に統計家は苦労させられるものである．さらに，データの量もさることながらそれを処理する計算機能は以前とはまるで異なってしまっている．統計的推測理論の中で統計量に対し正規分布による近似，χ^2-分布による近似といった形で処理されてきた部分をあらためて P-検定 の立場から「精密な形で」考えてみたい．ここで精密なという意味は検定にともなう過誤をキチンと評価できる手続きのことである．

本論文の目的は，一つは上記のような統計環境の変化を踏まえ，筆者が統計解析の方法として多用されてしかるべきと考える P-検定の仕組みと考え方を解説することにある．P-検定には少数例に対し有効であるという側面もある．その上で，カテゴリカルデータの分析に資するシフト係数などを用いた P-検定による解析法を示した．

本論文では第 2 節で通常の統計的仮説検定の考え方と進め方を 3 つの例をあげて説明する．第 3 節では P-検定の考え方と進め方を 2 つの場合について述べる．その上で，第 4 節では第 2 節であげた 3 つの検定法の場合が P-検定の立場からどう扱われるかを説明したい．さらに第 5 節ではカテゴリカルデータのシフトにかかわるシフト係数の利用が，P-検定によって有効に取り扱われることを示したが，これは新しい知見である．さらに従来から主に正規近似を主体として用いられてきた 2 つのノンパラメトリックな方法について P-検定の立場からの進め方を述べた．

2．仮説検定の考え方

統計的仮説検定といっても様々なタイプのものがあるが，本節では 1 標本の比率の検定，2 標本母平均の差の検定およびカテゴリカル度数についての χ^2-適合度検定を取りあげ，その考え方と方法を簡単に説明する．検定の考え方については主に松井 [13] の記述によっている．これらの検定法を後の節で P-検定 の立場からのアプローチで考え，比較，検討してみたい．

2．1　比率の検定（1 標本）

与えられた母集団 Π において，ある属性 C を持つものの占める割合（あるいは比率）が p で与えられ，その値が指定された値 p_0 と等しいといえるかどうかを判定したい．

母集団から得られた大きさ n のサンプルについて，属性 C の占める数を X とする．中心極限定理によって，$\hat{p} = X/n$ の分布は正規分布 $N(p,\, p(1-p)/n)$ によって近似される．このことを利用し，次のような手続きによって検定法を組み立てていく．

1. 有意水準 α を決める．

2. 帰無仮説 $H_0 : p = p_0$，と対立仮説を設定する．対立仮説は次のいずれか一つになる．

$H_1 : p > p_0$ または $H_1 : p < p_0$ （片側検定）

あるいは

$H_1 : p \neq p_0$ （両側検定）

3. 仮説 H_0 のもとでの $\hat{p}$ の分布を用い，

$$\text{統計量} \quad : \quad Z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \tag{1}$$

が，近似的に正規分布 $N(0,\, 1)$ にしたがうことを用いて検定する．

4. 結論を出す．判定は有意水準の値と対立仮説の設定のしかたで異なってくるが，統計量にデータを用いて計算した Z の実現値 z の値によってつぎのように帰無仮説を棄却するか否かを決定する．

棄却域はつぎのようになる．

片側検定：$z \geq z(2\alpha)$ （または $z \leq -z(2\alpha)$）

両側検定：$|z| \geq z(\alpha)$

上で使われている $z(\alpha)$ は標準正規分布 $N(0,\, 1)$ の両側 $100\alpha\%$ 点で，$\alpha = 0.05$ のとき $z(0.05) =$

1.96，$\alpha = 0.01$ のとき $z(0.01) = 2.576$ である．

例１： 大きさ $n = 100$（人）の無作為標本である意見に賛同しているものの数が 60 人であった．これに対する母集団における賛同者の割合は過半数を超えているとみてよいか？有意水準 5% で判定せよ．

帰無仮説と対立仮説を $H_0 : p = 0.5, H_1 : p > 0.5$ とする．

$\hat{p} = 60/100 = 0.60$ で，検定の式 (1) に代入すると

$$z = \frac{0.60 - 0.50}{\sqrt{0.5 \times (1 - 0.5)/100}} = 2.00.$$

仮説は右側で捨てることになるが，片側5 % 点は 1.645．したがって仮説 H_0 は棄却され，この結果から母集団における賛同者の割合は過半数を超えているとみてよい．

なお，$z = 2.00$ に対応する正規分布確率の値を P-値とよんでいる．この場合 P-値は 0.02275 となる．

２．２　母平均の差の検定（２標本，分散未知，同等）

２つの正規母集団 $N(\mu_x,\ \sigma_x^2)$, $N(\mu_y,\ \sigma_y^2)$ からとられた大きさ m および n の標本にもとづいて母平均の差についての検定を行なう．得られた標本平均をそれぞれ $\bar{X}$, $\bar{Y}$ とする．

それぞれの標本分布は

$$\bar{X} \sim N\left(\mu_x,\ \frac{\sigma_x^2}{m}\right), \quad \bar{Y} \sim N\left(\mu_y,\ \frac{\sigma_y^2}{n}\right)$$

である．$\sigma_x^2 = \sigma_y^2$ とし，推定量として不偏分散を用いる．検定の手続きはつぎの通りとなる．

1. 有意水準 α を決める．
2. 帰無仮説を $H_0 : \mu_x = \mu_y$ とし，対立仮説の設定はつぎのいずれか一つとする．

   $H_1 : \mu_x > \mu_y$ または $\mu_x < \mu_y$ （片側検定）

   あるいは

   $H_1 : \mu_x \neq \mu_y$ （両側検定）．
3. 分散 σ_x^2, σ_y^2 の不偏推定量を U_x^2, U_y^2 とし，平均の差の分散の推定量として

$$U^2 = \frac{(m-1)U_x^2 + (n-1)U_y^2}{m+n-2}$$

を用いる．仮説 H_0 の下で，検定統計量

$$\text{統計量} \quad : \quad t = \sqrt{\frac{mn}{m+n}}\frac{\bar{X} - \bar{Y}}{U} \qquad (2)$$

が，自由度 $m + n - 2$ の t-布にしたがうことを使って検定する．

4. 結論を出す．データから計算した値を上の統計量に用い，t の実現値を計算する．判定は自由度 $m + n - 2$ の t-分布の，両側あるいは片側 % 点を使って行う．

例２： ２つの正規母集団からつぎのようにデータが得られた（等分散 $\sigma_x^2 = \sigma_y^2$ を仮定）．

X : (8.82,　10.68,　10.02,　11.47,　9.01)
Y : (10.44,　12.08,　12.59,　11.12,　11.49,
　　9.71,　11.02,　9.55)

帰無仮説 $H_0 : \mu_x = \mu_y$ に対し，対立仮説を $H_1 : \mu_x < \mu_y$ として検定せよ．有意水準を 5% とする．

データから $\bar{x} = 10.0$, $\bar{y} = 11.0$ で，$U_x^2 = 1.2491$, $U_y^2 = 1.1471$，そして $U^2 = 1.18416$．したがって

$$t = \frac{10 - 11}{\sqrt{1.18416}} = -1.6120.$$

自由度 11 の t-分布の片側 5% 点は $t_{11}(0.05) = 2.201$ なので，仮説 H_0 は棄却されない．なお，P-値は 0.06763 である．

２．３　適合度検定（k カテゴリの場合）

大きさ n のデータをそれらの持つ属性によって B_1，B_2，...，B_k の k 個のカテゴリに分類したとき，それぞれ属性を持つものの度数を変数

$$(n_1,\ n_2,\ \ldots,\ n_k), \qquad \left(\sum_{i=1}^{k} n_i = n\right).$$

で表わす．

データが度数によってこのようにカテゴリカルに分類されているとき，それらのデータが背後に

ある分布法則にうまく適合しているかどうかを調べたいというのが適合度検定の目的である．たとえば，メンデルの分離の法則にかかわるメンデルによる実験では，交配によってそれぞれの形質をもつエンドウが 9 : 3 : 3 : 1，すなわち確率 9/16, 3/16, 3/16, 1/16 で得られるとしている（$k = 4$ である）．実験による検証では，観測による結果と期待確率から得られる期待度数との間の整合性があるかどうかをこの検定で調べることになる．

さて，適合度検定（goodness of fit test）に関わる表は表 1 のように表される．

表 1 適合度検定

	B_1	B_2	$\cdots$	B_k	計
観測度数	n_1	n_2	$\cdots$	n_k	n
期待度数	m_1	m_2	$\cdots$	m_k	n
期待確率	p_1	p_2	$\cdots$	p_k	1

ここで，$m_i = np_i$, $(i = 1, 2, \ldots, k)$ である．

この表については，つぎの多項分布が適用され，それぞれのカテゴリの期待度数 m_i は多項分布の各セルの平均に対応している．

$$M(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}, \quad p_i > 0, \ \sum_{i=1}^{k} n_i = n. \quad (3)$$

ところで，多項分布の正規近似の中で，観測度数と期待度数の適合性についてつぎの結果が得られている．

χ^2（カイ二乗）統計量

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - m_i)^2}{m_i} \quad (4)$$

は自由度 $k - 1$ の χ^2 分布にしたがう．

この統計量は観測度数が全体として期待度数に近い値を持っていればその値が小さくなる．そこで，設定した有意水準を 5% としたとき，仮説 H_0 :「観測度数は期待度数に適合している］を，対立仮説 H_1 :「そうとはいえない」に対して検定する．

実際に計算した値が自由度 $k - 1$ の χ^2-分布の 5% 点 $\chi^2_{k-1}(0.05)$ の値よりも大きければ，すなわち

$$\chi^2 \geq \chi^2_{k-1}(0.05)$$

のときに，仮説 H_0 を棄却する．

例 3 : 新製品の満足度調査で，製品のスタイルについての好感度を分類して表 2 の結果を得た．表内の記号は，D : よくない，C : あまりよくない，B : よい，A : 大変よい，である．

表 2 満足度調査

	D	C	B	A	計
回答数	21	19	25	35	100

好感度に関する反応は各カテゴリで等しいという仮説を検定せよ．

この場合の期待度数は $100/4 = 25$ である．(4) 式の χ^2-統計量を計算すると $\chi^2 = 6.08$ となる．自由度はカテゴリ数マイナス 1 なので，自由度 3 の上側 5% 点は $\chi^2_3(0.05) = 7.8147$．仮説は棄却されない．なお，この場合の P-値は 0.10778 である．

3．Permutation test の考え方

P-検定（permutation test）の考え方はノンパラメトリックな方法を中心に整理されてきたが，本節では順位和検定と連検定をテーマにこの考え方の基本を述べる．

3．1 順位和検定

薬の効果を調べるために，無作為に投薬したグループ（処置群）と投薬しないグループ（対照群）を各 3 人ずつ設定し，その結果にもとづいて効果を判定したいとする．効果の判定は医師により，効果の著しかった（良かった）ほうから $1, 2, 3, \ldots, 6$ までの順位をつけその結果にもとづいて効果の有無を判定するものとする．（Lehmann & D'Abrera [6] による例を改変）

処置群と対照群のそれぞれに 1 ～ 6 までの順位が割り振られる場合の数は $\binom{6}{3} = 20$ 通りある．順位 1 ～ 6 がそれぞれの群の対象者に割り振られたとし，これらすべてを記すと表 3 に示す通りと

なる．

ここで，表3のケース1では処置群の3名に効果の良かった順位 1, 2, 3 位が割り振られ，対照群の3名には効果の低い 4, 5, 6 位が割り振られたことになる．また，ケース 14 では効果を表わす順位 2, 4, 5 が処置群に，順位 1, 3, 6 が対照群に与えられたことになる．

表3 順位のすべての組合せ

	1	2	3	4	5	6
処置群	123	124	125	126	134	135
対照群	456	356	346	345	256	246

7	8	9	10	11	12	13
136	145	146	156	234	235	236
245	236	235	234	156	146	145

14	15	16	17	18	19	20
245	246	256	345	346	356	456
136	135	134	126	125	124	123

2つの群への効果の良さの指標として各群に与えられた順位の和を考えてみる．順位の和が小さければ群全体として効果があり，大きければ効果が小さかったと考えられる．順位の和は最小値が $1+2+3=6$，最大値が $4+5+6=15$ となっていて，20 組の処置群（対照群も同様）の和はこの間の値をとっている．

処置群での順位の和で考えると，上にあげたように最小値 6，最大値 15 で，全 20 ケースについての順位和の分布はつぎのように得られる．[*3]

表4 順位和の分布

順位和	6	7	8	9	10	11	
頻度	1	1	2	3	3	3	

順位和	12	13	14	15	計
頻度	3	2	1	1	20

ところで，薬の効果が全くなく，処置群でも対照群でも同じ結果が期待されるとしたらどうであろうか．一回の臨床検査の結果は表にあげた 20 通りの場合のいずれかがランダムに生ずると考えられ，それぞれの場合が起こる確率は 1/20 と考えられる．

ここで P-検定 ではつぎのように考える．帰無仮説として H_0：「処理の効果はない」とする．仮にケース1の場合が起こったとすると，このこと（処置群の順位の和が6で，対照群の順位の和が 15）の生起する確率は 1/20 であって，有意水準を 5% とすれば，このようなことが起こったということは H_0 を棄却して効果があったと見なすほうが自然であろうと判断するのである．[*4]

P-検定 は当該の問題に対し，与えられた条件のもとで起こりうるすべての場合を考え，その中で指標となる値およびその生起確率とあわせ判定を下すことになる．

3．2　連検定

ある事柄がランダムに起こっているか？逆にいうと，あるものの傾向が季節によって変動するか，あるいは周期性があるか？など，ある事柄の生起のランダム性について検定する方法に連検定がある．この検定法の考え方はつぎのように述べられる（松井 [13] 参照）．

反応が2分岐的な場合，同種のもののつながりを連（run）とよぶ．まず例をあげる．

例4：30 個の乱数列がつぎのように与えられている．これらの数字は奇数と偶数に関しランダムに出現しているとみなせるだろうか？（乱数は Hald [5] の乱数表から抽出.）

乱数を隣り合う数字の奇数と偶数との連なりによって「連」に分類していく．

0 | 5 9 7 | 0 | 3 | 4 4 | 9 | 6 4 | 9 | 8 | 7 1 3 | 0 | 1 | 6
9 5 | 6 | 1 9 | 2 | 7 3 | 0 | 3 1

この例では奇数の数が 18，偶数の数が 12，連の

[*3] Permutation test の名はこの例のように起こりうるすべての場合に対し，設定した統計量の値の分布を考えて判断することに由来している．この分布を後出のように permutation distribution とよんでいる．

[*4] 処置群の順位の和が 9（対照群の順位の和は 12）や処置群の順位の和が 10（対照群の順位の和は 11）の場合は「効果がない」という仮説の下ではむしろ普通に起こることと考える．

数（アンダーライン）が 20 である．

考え方

1 の数が $m = 3$，0 の数が $n = 2$，全体で $N = n + m$ 個の数の配列によって得られる全ての場合と対応する連の数は，つぎの 10 通りである．

表 5 連のパターンと連の数

1 1 1 0 0	2	1 1 0 1 0	4
1 1 0 0 1	3	1 0 1 1 0	4
1 0 1 0 1	5	1 0 0 1 1	3
0 1 1 1 0	3	0 1 1 0 1	4
0 1 0 1 1	4	0 0 1 1 1	2

これから「連の分布」がつぎのように得られる．

連の長さ	2	3	4	5
確率	2/10	3/10	4/10	1/10

この例では，(10101) という 0 と 1 のデータの配列の連の数は 5 である．このような事例の起こる確率は 0 と 1 の生起が独立でその起こる確率が同じであるとすれば，連の数という観点からみて，全 10 例中 1 例，すなわち 1/10 の確率で起こる．また，連の数が 2 というのは全 10 例中（パターンが 11100 および 00111 の）2 ケースであり，起こる確率は 2/10 となる．

このように見たときに，「連の数」が極端に多いのも，極端に少ないのもランダム性という仮説に反するというのがこの検定の考え方で，「連の分布」によって棄却域を決めることになる．

当該事例に対し設定した条件のもとで，ある事例の起こる相対頻度を起こりうるすべての事例に対して求め，その結果に基づいて帰無仮説（この場合は 0 と 1 の生起の等確率性ないしランダム性）についての判断を下す，というのが P-検定 の考え方である．

Good [4] は「P-検定 への五つの手順」としてつぎのようにあげている．上にあげた例と対比させると参考になるであろう．

1．問題の分析（仮説の設定など）
2．統計量の選択と判定基準
3．もとの観測値から統計量を計算
4．観測値を再配置（新しい配列で統計量を計算）
5．結論（全配列から得られた分布（パーミュテーション分布（P-分布）により判定）

4．仮説検定再考 － P-検定の立場から

本節では 2.1 節にあげた比率の検定，平均の差の検定および適合度検定を P-検定（permutation test）の立場から再考してみよう．

4．1　比率の検定

母集団が二項分布 $B(n,p)$ で，大きさ n の標本にもとづく個体数 x の割合 $\hat{p} = x/n$ にもとづき，つぎの仮説を検定する．

$H_0 : p = p_0$
$H_1 : p > p_0$ または $H_1 : p < p_0$ （片側検定）
あるいは
$H_1 : p \neq p_0$ （両側検定）

第 2.1 節で述べたように， n が大きい時には正規近似を用いて検定を行うことができる．ところが正規近似による方法には有意水準との関係で問題が存在する．この点を例で説明してみよう．

例 5 ： ある学生集団からランダムに選んだ学生 $n = 15$ 人に喫煙の有無を聞いた．「すう」が $m = 4$ 名，「すわない」が 11 名であった．「すう」に対する母集団比率を p としたとき，これをつぎの仮説の下で検定したい．

$H_0 : p = 1/2$ ， $H_1 : p < 1/2$

この場合，15 人を調査した結果「すう」とした人数 x のとりうるすべては 0 ～ 15 である．これらすべての x の値に対し，帰無仮説の下でこれらの値がとられる確率がつぎの二項分布によって計算される．

$$\Pr(x) = \frac{15!}{x!(15-x)!}\left(\frac{1}{2}\right)^{15}, \quad x = 0, 1, 2, \ldots, 15. \quad (5)$$

この確率分布（permutation distribution）は左右対称形となっているが，与えられた x に対する確率とその累積確率の一部を記すと表 6 のようになる．

「すう」が 4 名なので，この場合 P-値は 0.0592

である．ところで，有意水準を 5% としたとき，この分布は離散分布なので，その値に明確に対応する点は存在しないが，仮説は棄却されない．

表 6 二項分布確率 $B(15, 1/2)$ とその累積和

x	0	1	2	3
$\Pr(x)$	0.0000	0.0005	0.0032	0.0139
$\sum \Pr(x)$	0.0000	0.0005	0.0037	0.0176
x	4	5	6	7
$\Pr(x)$	0.0417	0.0916	0.1527	0.1964
$\sum \Pr(x)$	0.0592	0.1509	0.3036	0.5000

2.1 節にあげた正規近似を使うと，

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{4 - 7.5}{\sqrt{15/4}} = -1.8074$$

これは -1.645 （右側 5% 点）より小さいので仮説は棄却される．これによる P-値は 0.03535 である．全事例を計算することによって得られた正確な値（結果）と正規近似による結果が異なることとなった．

二項分布の正規近似は n が十分大きいときとしている．この場合（$n = 15$）の正規分布への近似はかなり良好であるが，二項分布の表に合わせて確率を計算すると表 7 が得られる．また，図 1 に二項確率の分布と対応する正規分布を示した．表には二項分布と正規分布の累積確率をあげ，正規分布では切点を x とした場合と $x + 0.5$（補正）した場合をあげ，どの程度の違いがあるかを示した．

結果はかなりの差異を示しており，このために連続性の補正を行う場合がある．これは近似の x に対し $x \pm 0.5$ をとり離散変量と連続分布である正規分布の区間幅の調整を行っている．

いずれにしても，両分布間の確率の乖離は生じているわけで，この場合 5%，1% といった形で有意水準を設定し検定するのではなく結果に対する P-値で処理するほうが賢明である．

例 6 ： 例 1，賛同者の割合

計算手段の向上で例 1 にあげた賛同者の割合の例のように n の値がかなり大きい場合でも二項分布による厳密な計算は容易にできるようになってい

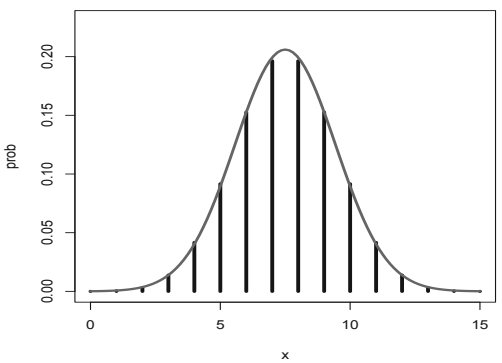

図 1　二項分布の正規近似，$n = 15, p = 0.5$

る．二項分布によると，60 回以上表の正確な値は

$$\Pr(X \geq 60) = 0.028444$$

である．正規近似による場合は

$$z = \frac{x - np}{\sqrt{np(1-p)}} = \frac{60 - 50}{\sqrt{25}} = 2.0$$
$$\rightarrow \quad \Pr(Z \geq 2.0) = 0.022750$$

正規近似で，連続性の補正を行った場合は

$$z = \frac{x - 0.5 - np}{\sqrt{np(1-p)}} = \frac{59.5 - 50}{\sqrt{25}} = 1.9$$
$$\rightarrow \quad \Pr(Z \geq 1.9) = 0.028717$$

となる．このように n が大きいときには正確な値と正規近似の値とはかなり近くなる．

4．2　母平均の差の検定

正規分布 $N(\mu_x,\ \sigma_x^2)$，$N(\mu_y,\ \sigma_y^2)$ にしたがう大きさ m および大きさ n の標本

$$\mathrm{X} : (X_1, X_2, \ldots, X_m), \quad \mathrm{Y} : (Y_1, Y_2, \ldots, Y_n)$$

にもとづいて母平均の差の検定を行いたい．このとき，分散 σ_x^2，σ_y^2 が未知で等分散が仮定できる場合には通常 2.2 節にあげた統計量 t を用いて検定する．これを P-検定 の立場から検討してみよう．

いま，等分散 $\sigma_x^2 = \sigma_y^2$，帰無仮説として H_0 : $\mu_x = \mu_y$ とすると仮説が正しいとしたときに得ら

表 7 二項確率と正規近似（切点との関係）

x	0	1	2	3	4	5	6	7
二項（累積）	0.0000	0.0005	0.0037	0.0176	0.0592	0.1509	0.3036	0.5000
正規 $\leq x$	0.0001	0.0004	0.0023	0.0101	0.0354	0.0984	0.2193	0.3981
正規 $\leq x+0.5$	0.0002	0.0010	0.0049	0.0194	0.0607	0.1508	0.3028	0.5000
x	8	9	10	11	12	13	14	15
二項（累積）	0.6964	0.8491	0.9408	0.9824	0.9963	0.9995	1.0000	1.0000
正規 $\leq x$	0.6019	0.7807	0.9016	0.9646	0.9899	0.9977	0.9996	0.9999
正規 $\leq x+0.5$	0.6972	0.8492	0.9393	0.9806	0.9951	0.9990	0.9998	1.0000

れた X と Y を合併した大きさ $m+n$ の標本は，同一母集団からのランダムサンプルと考えられる．このうちの大きさ m の標本にもとづく実現値が実際に得られた X に関するデータである．

ここで，平均の差をつぎのように書くことができる．

$$\bar{X}-\bar{Y}=\frac{m+n}{mn}\sum_{i=1}^{m}X_i-\frac{1}{n}\left(\sum_{i=1}^{m}X_i+\sum_{j=1}^{n}Y_j\right) \tag{6}$$

$m+n$ 個の中から m 個を抽出する．このような抽出の仕方は $\binom{m+n}{n}$ 組存在する．そのすべての組の抽出に際し $\sum_{i=1}^{m}X_i+\sum_{j=1}^{n}Y_j$ は（全体の値に対応しているので）固定されている．すなわち，差 $\bar{X}-\bar{Y}$ を考えることは

$$(X_1, X_2, \ldots, X_m, Y_1, Y_2, \ldots, Y_n)$$

の $m+n$ 個の中から無作為に m 個を抽出したときの，その結果の平均 $\bar{X}$ の挙動にのみ関係する．このとき，考えられるすべての組によって作られる $\bar{X}$ の分布がこの場合に得られた permutation distribution（以後，P-分布 と表わす）である．[*5]

この分布に，実際に得られたデータの平均 $\bar{X}$ の値を対決させることによってこの場合の P-値が得られる—すなわち検定が可能になる．

例 7 ： 2.2 節の例

[*5] (6) 式で差 $\bar{X}-\bar{Y}$ の P-分布を得ることにしても結論は同じとなる．$\bar{X}$ に限定した方が計算は楽になる．

大きさ $m=5$ と $n=8$ のデータがつぎのように得られている．

X : (8.82, 10.68, 10.02, 11.47, 9.01)
Y : (10.44, 12.08, 12.59, 11.12, 11.49, 9.71, 11.02, 9.55)

これら X と Y を合併した中から大きさ 5 のサンプルを抽出する．この場合の数は $\binom{13}{5}=1,287$ 通りある．これらによって作られる平均値の全体が P-分布 を構成しているが，分布のヒストグラムは図 2 のようになっている．

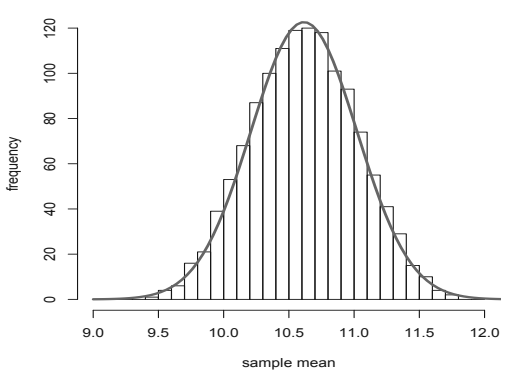

図 2 $\bar{X}$ の P-分布

この P-分布 の平均は 10.6154, 分散は 0.16516 であった（不偏分散は 0.16528）．実際に得られたデータ X についての平均は上に示した通り 10.0 なので，この値に等しくなる確率を P-分布 の中から探

すことになる．分布は離散分布で，結果をパソコン上で値の大きさの順に並べ替えて整理すると，分布の小さい値の方から 87 番目がデータの平均 10.0 と一致した．この確率（P-値）は $87/1287 = 0.06760$ である．したがって，有意水準を 5% とすると母平均が等しいという仮説は棄却されない．

$\bar{X}$ の P-分布 を $\bar{X}$ の平均と分散を持つ正規分布で近似したのが図 2 上の近似曲線である．この正規分布によって P-値を求めると

$$Z = \frac{10 - 10.6154}{\sqrt{0.3823}} = -1.5143$$
$$\Longrightarrow \quad \text{P-値は } 0.06498$$

となる．なお，通常の平均の差による t-検定の結果は 2.2 節にあげたが，P-値は 0.06763 であった．

4．3　適合度検定

χ^2-適合度検定の一般の形は 2.3 節で述べた．ここでは P-検定 の立場からのアプローチを示してみよう．カテゴリ数が $k = 4$，度数総計が $n = 100$ の場合を例に説明する．他の k，n についても同様に進められるが，値が大きくなると後で述べるように全事例を導出するのが大変で計算上の負担が生ずる．

一般に k 個のカテゴリに度数を $(n_1, n_2, \ldots, n_k)$，$n_1 + n_2 + \cdots + n_k = n$ となるように配分する方法は

$$N(k, n) = \binom{n+k-1}{k-1} \tag{7}$$

通りある．これから説明する例では $k = 4, n = 100$ なので

$$(n_1, n_2, n_3, n_4), \quad n_1 + n_2 + n_3 + n_4 = n$$

と配分される場合の数は $N(4, 100) = \binom{103}{3} = 176,851$ 通りとなる．例にあげるにはかなり大きな数だが，計算上はこの程度でも問題はない．カテゴリ確率と合わせたこの時のカテゴリカルモデルは表 1 で $k = 4$ の場合である．また，χ^2-統計量は (4) 式で与えられる．ここで，帰無仮説を $H_0 : p_1 = p_2 = p_3 = p_4 (= 1/4)$ とおくと[*6]

$$\chi^2 = \frac{4}{n} \sum_{i=1}^{4} \left(n_i - \frac{n}{4} \right)^2 \tag{8}$$

となる．

ここで $N(4, 100) = 176,851$ 通りのすべてのカテゴリ度数 (n_1, n_2, n_3, n_4) について上記の χ^2 の値を計算し得られるのが，仮説 H_0 のもとでの P-分布 である．すなわちつぎのパターンとなっている．

カテゴリ度数	χ^2 統計量	カテゴリ度数の確率
(n_1, n_2, n_3, n_4)	χ^2 の値	$\Pr(n_1, n_2, n_3, n_4)$

この対応によって統計量とそれに対する確率を表わす P-分布が計算される．$\Pr(n_1, n_2, n_3, n_4)$ は配列 (n_1, n_2, n_3, n_4) に対する多項分布確率で，一般形は式 (3) である．

この結果（P-分布）を図示すると図 3 のように得られる．

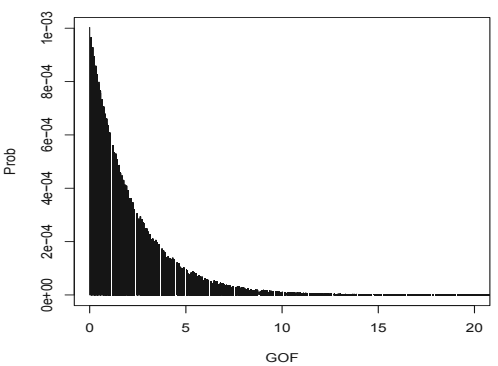

図 3　χ^2-値の P-分布（$k = 4, n = 100$）

通常は 2.3 節で説明したようにこれを自由度 $k - 1 = 3$ の χ^2-分布で近似して棄却限界値を求め検定を行っている．

この P-分布 で得られる確率と自由度 2 の χ^2-分布から得られる上側パーセント値を与えたのが表 8

[*6] この例では帰無仮説を等確率としたが，一般には $H_0 : p_1 = p_1^0,\ p_2 = p_2^0,\ p_3 = p_3^0,\ p_4 = p_4^0,\ p_1^0 + p_2^0 + p_3^0 + p_4^0 = 1,$ のように任意に設定できる．

である．P-分布 は 176,851 個の離散点に対して与えられた確率の分布で 4.1 節で説明したように有意水準と等しくなるような統計量の値を必ずしも見いだせない．そこで，表には内輪の値をとり，これを実際の値の欄にあげている．

表 8　P-分布と χ^2 の限界点

%	統計量の限界点	実際の確率	χ^2 の近似値
0.5	12.80	0.00489	12.838
1.0	11.44	0.00966	11.345
2.5	9.44	0.02309	9.348
5.0	7.84	0.04932	7.815

χ^2-分布への近似はかなり良いことがうかがえるが，多少の乖離はみられる．

P-検定では得られる χ^2 の値に対応する棄却限界値（P-値）を仮説の下での P-分布 から求め，判定を下すことになる．この場合も結果の判断には有意水準よりも P-値を用いることが望ましいといえる．

例８： 2.3 節の例 3 に対し P-分布 から得られた結果

全体で $N(4,100) = 176,851$ 通りの場合について χ^2-値と四項分布による生起確率を計算することによってこの場合の P-分布 が得られる．例 3 の結果から $\chi^2 = 6.08$ で，P-分布 上でこの値に等しい確率を見つける．P-分布 上にはこの値に等しいカテゴリ度数の場合が 72 例あり，上側確率（P-値）が最小のものが 0.103622，最大のものが 0.107107，その中間の値が 0.105120 であった．なお χ^2-分布を用いた近似では上側確率は $\chi^2_2(6.08) = 0.107782$ である．

５．P-検定の実際

本節では P-検定（permutation test）の実際を，5.1，5.2 節ではシフト係数を例に，5.3 節では順位和検定，5.4 節では連検定を例に説明したい．

５．１　シフト係数の場合

シフト係数（RCN，relative category number）は表 1 にあげたようなカテゴリカルデータ

$$A(k) = (n_1, n_2, \ldots, n_k), \qquad \sum_{i=1}^{k} n_i = n. \quad (9)$$

について，カテゴリ度数のシフトの状況（度数が右側のカテゴリで大きいか，左側のカテゴリで大きいか）を示す統計量で，以下カテゴリ $A(k)$ のシフト係数を $RCN(A(k))$ と表わすことにする．シフト係数は -1 から $+1$ までの間の値をとり，相関係数と似た役割を果たしている．シフト係数はカテゴリ確率に関する仮説に対し，データの適合性の判定にも用いられる．詳しくは Matsui [8]，松井 [14] を参照されたい．

表 1 で，全体で大きさ n のデータを k 個のカテゴリに対し $n_1 + n_2 + \cdots + n_k = n$ のように配分する方法は (7) 式の $N(k,n)$ 通り存在する．このすべての $(n_1, n_2, \ldots, n_k)$ の組に対し定義されるのがシフト係数 RCN である．シフト係数はカテゴリ数 k と度数配列 $A(k)$ によって計算される統計量だが，$k \geq 4$ の場合には式の表現が大部なもの（項数が多い）になるので，数値計算上は漸化式を用いて求めることが勧められる（本稿でも松井 [14] にあげた R 言語によるスクリプトを用いて RCN の計算を行っている．）

シフト係数は，カテゴリカル度数の左右へのシフトを $-1 \sim +1$ の範囲の係数としてとらえているが，カテゴリ確率に関して与えられた確率モデルの下でその分布が規定されるので，この後の例 9 で示す等確率の仮説の下でも，2.3 節にあげたメンデルの分離の法則を仮説値とするような場合にも P-検定の考え方を利用して当該の仮説に対し検定を行うことができる．

シフト係数を用いた P-検定は適合度検定について 4.3 節で説明したことと基本的に同じである．4.3 節で統計量として χ^2 を用いたところに $k = 4$ の場合のシフト係数を適用し，後の手続きも同様に進めていく．$k = 4$，$n = 100$ の場合のシフト係数による P-分布 は帰無仮説として $H_0 : p_1 = p_2 = p_3 = p_4$ としたとき図 4 のように

得られる．シフト係数も n が大きいとき正規近似ができる．

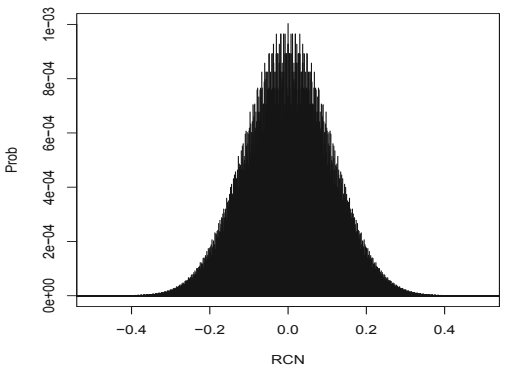

図 4　シフト係数の P-分布（$k = 4, n = 100$）

例 9 ： 2.3 節，例 3

この例ではカテゴリ数 $k = 4$ で，カテゴリ度数が (21, 19, 25, 35) となっている．この度数配列に対するシフト係数は 0.2191 となる．考えられるすべての度数配列の数は $N(4, 100) = \binom{103}{3} = 176,851$ であり，P-分布 は図 4 である．この分布からシフト係数 0.2191 に対応する上側確率は 0.021402 と得られるが，これが P-値である．すなわち，等確率の仮説 H_0 に対し，上のデータから 5% 水準で考えて，右への有意なシフトを認めることができる（等確率であるという仮説は棄却される）．

5．2　カテゴリカル・リスポンスへの応用

本節では，授業評価の比較を扱った事例をもとにシフト係数による P-検定の例を説明する．

ある科目の授業の理解度に関する授業評価で女性と男性それぞれ 10 人の評価結果が表 9 のように得られた．なお，評価の数字はつぎのようになっている．

1 ：まったく理解できない
2 ：よく理解できない
3 ：どちらともいえない
4 ：ある程度理解できる
5 ：十分に理解できる

表 9　授業評価と人数

評価者	1	2	3	4	5	計
女性	0	1	3	4	2	10
男性	0	1	3	5	1	10

この例では男女による評価結果の差はかなり微妙といえ統計的な判断を下すのは難しいように思われる．この問題にシフト係数を適用することによってカテゴリカル度数のシフトの状況（この場合は理解度の高い評価のへの傾向性）や，カテゴリ度数の一様性からのズレを判定することができる．シフト係数を用いた P-検定 による手続きはつぎの通りである．

カテゴリ数 $k = 5$ で $n = 10$ なので，可能な配列総数は $N(5, 10) = 1001$ である．このすべての配列についてシフト係数（RCN）の値と帰無仮説 $H_0 : p_1 = p_2 = p_3 = p_4 = p_5$ の下で $k = 5$ の場合の多項分布確率を計算する．これによって表 10 の形式の結果が得られる．表 10 は作成される 1001 の配列全体の表の一部で，シフト係数の値について降順に，累積確率をあわせ製表してある．[*7][*8]

$k = 5$ のシフト係数の計算から，表 9 にあげた女性による評価の配列 $A = (0, 1, 3, 4, 2)$ についてのシフト係数は $RCN(A) = 0.5320$，男性による評価の配列 $B = (0, 1, 3, 5, 1)$ については $RCN(B) = 0.3830$ であり，配列 A の右へのシフト（理解度への高い評価）は大きい．さらに，仮説 $H_0 : p_1 = p_2 = p_3 = p_4 = p_5$ としたときのシフト係数についての P-分布は図 5 のように得られるが（表 10 はこの一部分である），この分布から，

$$\Pr(RCA(A) \geq 0.5320) = 0.04367$$

[*7] P-検定 を説明したこれまでの節でも表 10 と類似の表をパソコン上に $N(k, n)$ 通りのすべてにわたって作成し，これから P-分布 を作成している．これまでの例では表が大部になるので示さなかったが，手続き上はここに示すように進める．

[*8] この例では各カテゴリの確率が等しいという仮説を帰無仮説として P-分布を求め，有意性の判定を行っている．たとえば，帰無仮説を「H_0 : すべてのカテゴリ配列は同様に確からしい」とすると異なる P-分布が得られる．その仮説の下では女性データの P-値は $P = 0.1259$，男性データの P-値は $P = 0.2018$ である．

表 10　P-分布の一部（$k = 5,\ n = 10$）

配列順	カテゴリ度数					RCN	多項分布確率	累積確率
124	1	0	3	2	4	0.5340	0.00129	0.04230
125	0	0	6	1	3	0.5320	0.00009	0.04359
126	0	1	3	4	2	0.5320	0.00129	0.04367
127	0	1	5	0	4	0.5310	0.00013	0.04496
128	0	0	5	3	2	0.5270	0.00026	0.04509
200	1	0	3	4	2	0.3880	0.00129	0.10585
201	0	3	3	1	3	0.3870	0.00172	0.10714
202	0	1	3	5	1	0.3830	0.00052	0.10886
203	0	1	6	0	3	0.3810	0.00009	0.10938
204	2	0	2	1	5	0.3810	0.00077	0.10946

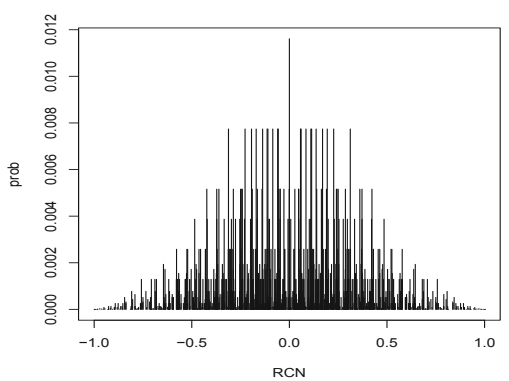

図 5　シフト係数の P-分布（$k = 5, n = 10$）

すなわち，配列 A の上側 P-値は 0.04367 である．同様に，配列 B の P-値として 0.1089 が得られる．この結果から，有意水準を 5% としたとき帰無仮説 H_0 に対し，配列 A は有意に右にシフトしているといえるが，配列 B についてはそのようには言えないということが分かる．

5．3　順位和検定の場合

順位和検定は比較したい 2 つの群に与えられた順位をもとに，帰無仮説 H_0 :「2 つの群は同じ分布からとられた」を検定する．対立仮説の表現にはいろいろあるが，「2 つの分布の中央値あるいは位置母数にズレがある」である．

2 群の観測値があって，X 群からの大きさ m の観測値，Y 群からの大きさ n の観測値がある指標によって 1 から $m + n$ までの順位で与えられた場合を考える．それぞれに与えられた順位の和が検定のために用いられる順位和を構成する．

もとのデータが順位でない場合にも有効で，たとえば，例 2 で与えた X，Y のデータを合併し，それぞれの大きさにしたがってつぎのように順位を割り振ったものでもよい．

X の順位 $\Rightarrow$ (1, 7, 5, 10, 2)　順位和 25

Y の順位 $\Rightarrow$ (6, 12, 13, 9, 11, 4, 8, 3)　順位和 66

この場合，例 2 では X，Y のそれぞれが正規分布からのデータであるとし，検定統計量もその仮定の上で採用しているが，順位和検定は分布形が不明の場合のノンパラメトリックな方法として使えることを示している．

X と Y に与えられた順位データの和をそれぞれ R_x，R_y とすると，

$$R_x + R_y = \frac{1}{2}(m + n)(m + n + 1) \qquad (10)$$

である．それぞれの平均を

$$\bar{R}_x = \frac{1}{m}R_x, \quad \bar{R}_y = \frac{1}{n}R_y \qquad (11)$$

とすると，4.2 節 (6) 式の場合と同様に

$$\bar{R}_x - \bar{R}_y = \frac{m+n}{mn}R_x - \frac{1}{n}(R_x + R_y) \qquad (12)$$

という関係があり，$R_x + R_y$ は全順位の和で一定なので，順位の平均の差による検定には $\bar{R}_x$ を使えばよいことが分かる．また，順位和の差で比較を考えるにしても

$$R_x - R_y = 2R_x - (R_x + R_y) \qquad (13)$$

となるので，順位和 R_x のみを考えればよい．

P-分布 を得るための手続きは 4.2 節で説明したものと同じである．1 から $m+n$ までの数字の中から m 個を抽出し和（順位和）をつくる．これを場合の数 $\binom{m+n}{m}$ のすべてについて実行し順位和（あるいはその平均）の P-分布 を得る．これを実際に得られたデータに対決させて P-値を求める．

この P-分布 を得るプロセスは m と n の値が大きい場合には面倒であるが，精密な結果を得るための方法としては優れている．

m, n が十分大きい（一般に $m, n > 10$）の場合には正規近似によるつぎの統計量が使える．仮説の下で，順位和の平均と分散は（Lehmann & D'Abrera [6]）

$$E(R_x) = \frac{1}{2}m(m+n+1), \qquad (14)$$

$$V(R_x) = \frac{1}{12}mn(m+n+1) \qquad (15)$$

なので，

$$Z = \frac{R_x - E(R_x)}{\sqrt{V(R_x)}} \sim N(0,\ 1) \qquad (16)$$

あるいは連続性の補正を用いて（上側確率の場合はマイナス，下側確率の場合はプラスをとる）

$$Z = \frac{R_x \pm 0.5 - E(R_x)}{\sqrt{V(R_x)}} \sim N(0,\ 1). \qquad (17)$$

例 10： 2.2 節，例 2 を順位化したデータ．

例 2 を順位化したデータは上にあげておいた．$m = 5$，$n = 8$ で，$R_x = 25$ である．

この場合に考えられるすべての場合は 13 個の中から 5 個を取り出す場合の数で 1,287 通りである．このすべての場合について X の順位の順位和を求めることで R_x の P-分布 が得られる．ヒストグラム化した概要を図 6 に示した．この P-分布 の平均は 35，分散は 46.6667 である．

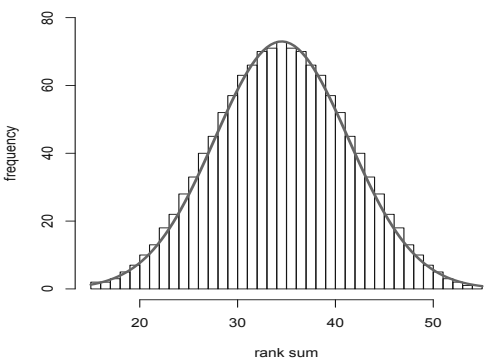

図 6　順位和の P-分布

この分布にデータから得られる順位和 $R_x = 25$ を対決させると，度数が同じになるケースが複数（28 個）あるので，下側確率（P-値）は最小，中間，最大として 0.0645, 0.0750, 0.0855 となる．

m，n の値は大きくないが，正規近似をあえて行うと $E(R_x) = 35$，$V(R_x) = 46.6666$ で，(16) 式では $z = (25-35)/\sqrt{46.6666} = -1.4639$．これから P-値は 0.07162 となる．(17) 式では $z = (25 + 0.5 - 35)/\sqrt{46.6666} = -1.3907$．P-値は 0.08216 である．

5．4　連検定の場合

3.2 節で連検定を使って P-検定 の考え方を示したが，連検定は 2 種類の識別子の混在した列について，そのランダム性ないし傾向性の有無を検定する方法である．

与えられたデータについて，2 種類の識別子の数がそれぞれ m, n 個，連の数が R 個であったとする．このとき P-検定 の考えにしたがって識別子の並べ方のすべてを考え連の数の P-検定 を求めることは m, n の値が大きいときには簡単でない．そこで一般にはつぎの正規近似による方法を使う．

m, n の値が大きいとき，2 つの識別子が等確率で出現するという仮説の下で，連 R はつぎの平均と分散を持つ正規分布によって近似できる（Siegel and Castellan [11]）．

$$E(R) = 1 + \frac{2nm}{m+n} \tag{18}$$

$$V(R) = \frac{2mn(2mn-m-n)}{(m+n)^2(m+n-1)} \tag{19}$$

これから

$$Z = \frac{R - E(R)}{\sqrt{V(R)}} \sim N(0,\ 1) \tag{20}$$

例 11： 3.2 節，例 4 の乱数列の例では奇数 $m = 18$，偶数 $n = 12$，連の数 $R = 20$ である．この乱数列はランダム性を保持していると見なせるだろうか．

正規近似を用いると，上の式から $E(R) = 15.4$, $V(R) = 6.65379$．これから

$$z = \frac{20 - 15.4}{\sqrt{6.65379}} = 1.78329.$$

正規分布の片側 5% 点が 1.645 なので，この乱数列はランダム性を保持しているとは見なせない．なお，P-値は 0.03727 である．

なお，(19) 式で連続性の補正を行うと $z = (20 + 0.5 - 15.4)/\sqrt{6.65379} = 1.977135$ で，P-値は 0.02401 となる．

ところで，連の数についてはつぎの結果が知られている（Wilks [12] 参照）．2 つの識別子を「0」と「1」とし，m 個と n 個の 0-1 の列が確率 p で「1」が起こり，確率 $1-p$ で「1」が起こるベルヌーイ試行の列であるとする．連の数 R の分布は p に依存するが，「0」の数が m，「1」の数が n という条件のもとで R の分布はつぎのように与えられる．

$$\Pr(R = 2k) = \frac{2\binom{m-1}{k-1}\binom{n-1}{k-1}}{\binom{m+n}{n}} \tag{21}$$

$$\Pr(R = 2k+1) = \frac{\binom{m-1}{k-1}\binom{n-1}{k} + \binom{m-1}{k}\binom{n-1}{k-1}}{\binom{m+n}{n}} \tag{22}$$

したがって，この確率分布の計算を準備しておけば (20) 式の正規近似によらずに精密な検定が可能となる．

図 7 に連の分布とそれに正規分布を当てはめた例を示しておいた．

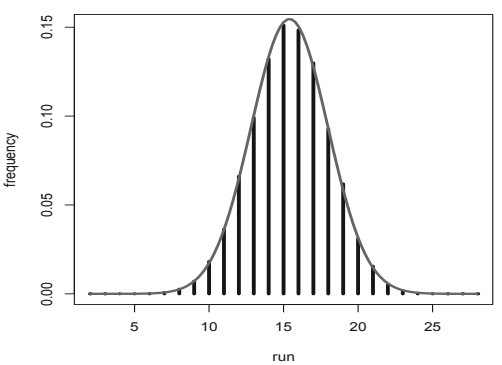

図 7　連の分布と正規分布

例 12： 例 4 および例 11 の乱数列の例で，$m = 18$, $n = 12$ の場合の P-分布 を式 (21), (22) によって求め，同じランダム性の仮説を調べてみる．この場合連の総数は $\binom{m+n}{m} = \binom{30}{18} = 86,493,225$ 通りである．データから得られた連の数は $R = 20$ であるが，上の連の確率分布の中で，連の数が 2 から 20 までの事例は 84,541,795 通りと計算される．これから $\Pr(R \le 20) = 0.977438$，すなわち $\Pr(R \ge 21) = 0.02256$ となる．したがって精密検定の結果でもランダム性の仮説は棄却される．

連検定の場合には上記のように連の確率分布が得られているので，これを用いると P-値が正確に計算される．サンプルサイズ m, n がかなり大きいときには正規近似も有効である（ただし，この場合連の数を求める自動化も必要となる）．

6．おわりに

本節ではこれまでに記述してない点を補遺的にあげておきたい．

Permutation test（P-検定）と類似の方法に randomization test がある．これは Edgington [1] によれば無作為実験において処理効果についての帰無仮説を無作為化（無作為割り付け）によって検定する，すなわち randomization model に対応するタイプの P-検定であるといっている．しかし，Ernst [2] によれば population model に適用された場合も含めこの形の検定は一般的に permutation test とよばれることが多い．本稿では呼称よりも検定の考え方と方法をテーマにしているので，より広い意味で使われている permutation test を採用している．

なお，P-検定の説明と直接関係するわけではないが，上記のこととの関連でいうと，本稿にあげた例では randomization model の例としては 3.1，5.3 節の順位和検定，3.2，5.4 節の連検定があげられ，population model の例としては 2.1，4.1 節の比率の検定，2.2，4.2 節の母平均の差の検定があげられる．また 2.3，4.3 節の適合度検定，5.1 節のシフト係数の例は hybrid model（Ludbrook & Dudley [7]）といえる．

はじめにでも述べたように，Fisher [3] は C. ダーウィンのデータに対し，平均の差にかかわる P-検定を考え，32,768 通りの並べかえによって P-分布を得ている．これらはすべて手計算でなされたわけだが，難しいというよりは単調な計算量は確かに煩雑である．ふりかえって，本論文の中で扱った例では全事例配列の数がかなり大きい場合も含まれているが，計算上の支障は特に感じられなかった．適合度検定やシフト係数の例で扱った異なる配列の数は $N(4,100) = 176,851$ 通り，差の検定と順位和検定では 1,287 通りである．5.4 節の連検定では場合の数は多いが理論的な確率分布が分かっているので問題はない．計算上のハードルはかなり低くなっているといえる．この意味においても，むしろ P-検定にむけたアルゴリズムやプログラム開発があってよいし，もっと多用されるべきであろう．

5.1，5.2 節のシフト係数による方法はカテゴリカルデータの左右方向へのシフトを測るものとして考案された．事例（例 3，例 9）に与えたように，等確率性に関する適合度検定では有意性が見られなかった結果に対し，同じ仮説からシフト係数による方法ではカテゴリの右へのシフトが認められる結果が得られている．また，シフト係数は（一般には P-検定は）5.2 節のような少数例に対しても有効で，このようなケースに対しては伝統的な諸方法ではなかなかうまく対応できなかったわけである．

最後に，P-検定は検定の方法としてデータの数（サンプルサイズ）が少数の事例に対しても有効である点をもう一度強調して筆をおきたい．

参考文献

[1] Edgington, E.S., Randomization Tests, Third Ed., Marcel Dekker, Inc., 1995.

[2] Ernst, M.D., Permutation methods: A basis for exact inference, Statistical Science, Vol.19, No4, 2004.

[3] Fisher, R.A., The Design of Experiments, (first published in 1935), Hafner Pub. Co., 1971.

[4] Good, P., Permutation Tests, Second Edition, Springer, 2000.

[5] Hald, A., Statistical Tables and Formulas, J. Wiley & Sons, 1967.

[6] Lehmann, E.L., D'Abrera, H.J.M., Nonparametrics, Holden-Day, Inc., 1975.

[7] Ludbrook, J., Dudley, H., Why permutation tests are superior to t and F tests in biomedical research, American Statistician, Vol.52, No.2, 1998.

[8] Matsui, T., Testing a shift of categorical response probabilities for the associated probability model, 獨協経済, 第 78 号, 2004.

[9] Moore, T.L., Condit, V. B., Using permutation tests to study infant handling by female

baboons, Stats, No.31, 2001.
[10] Salsburg, D., The Lady Tasting Tea, W.H.Freeman & Co., 2001.
[11] Siegel, S., Castellan,Jr., N.J., Nonparametric Statistics for the Behavioral Science, Second Edition, McGraw-Hill Co., 1988.
[12] Wilks, S.S., Mathematical Statistics, J.Wiley & Sons, 1962.
[13] 松井　敬,「統計的推測」, 共立出版, 2012.
[14] 松井　敬, カテゴリカルデータの分析―シフト係数とその応用，情報学研究，Vol.3, 2014.