

● 講座/臨床研究のための統計学

I. 対応のない場合の2標本の比較*

高木 廣文**

はじめに

臨床研究において、治療法や薬剤の効果を比較検討する場合、統計学の手法を用いる必要がある。しかし、臨床に携わる大多数の人にとって、統計学は面倒で煩わしいものに違いない。その理由の一つとして、統計学の手法が多数あるため、実際の研究にどれを適用すればよいか、判断に迷うという点があげられよう。また、適当な手法を選んで、その選択は正しかつたのか、結果の解釈に誤りはないのかなど、若干の不安が残ることもあるだろう。

上記のような問題点を考慮し、ここでは従来とやや異なる方法により、統計学の解説を行うことにした。解説の手順として、まず例題を示し、それに対する計算の手順と結果の評価を行い、その後、に用いた手法の理論と適用上の問題点などについて、簡単な説明を加えることにした。なお、各検定で用いる統計数値表は、統計学の成書には必ず掲載されているので、ここでは紙面の都合もあり、割愛することにした。読者は、それらの数値表を必要に応じて参照して欲しい。

実際にこれから解説していく手法は、以下のよう内容である。すなわち、

- 1) 対応のない場合の2標本の比較
- 2) 対応のある場合の2標本の比較
- 3) クロス表の検定
- 4) 特殊な形式のデータの検定
- 5) 分散分析とそれに関する検定
- 6) 生命表解析について

などである。上記の分類の表題は、その中に含まれる手法にとつて、必ずしも最適でないかもしれない。したがって、各回の表題とその内容に、一致しない部分が生ずるかもしれない。それらの点に関して、読者が混乱しないことを願う次第です。

対応のない場合の2標本の比較

1. 独立な2標本の母平均値の差の検定

例題 1・1

急性膀胱炎患者を無作為に2群に分け、一方に新薬Aを投与し、他方に旧薬Bを投与した。各群の標本数は50例、症状消失までの平均日数と標準偏差(不偏分散の平方根)は、A群は3.1(日)と1.2、B群は5.3と0.8であった。この成績から、A薬はB薬より優れているといえるだろうか。

【検定法 1・1】

ここでは、まず2群の母分散は等しいものと仮定して、母平均値の差の検定を行う。

後述する(1.3)式に各数値を代入し、

$$t_0 = \frac{|3.1 - 5.3|}{\sqrt{\frac{(50-1) \times 1.2^2 + (50-1) \times 0.8^2}{50+50-2} \times \left(\frac{1}{50} + \frac{1}{50}\right)}} \div 10.786$$

を求める。t₀の自由度は(50+50-2)=98となる。t分布表から自由度98の両側5%点、1%点、0.1%点を求めると、1.986、2.629および3.397である(数値表に対応する自由度が掲載されていない

* Statistics in Clinical Studies: I. Methods of the Comparison for the Independent Samples

** 聖路加看護大学 Hirofumi Takagi: St. Luke's College of Nursing

場合、補間法により求めるか、その自由度より小さいもので最も近い自由度に対する数値を用いればよい。これらの数値と t_0 の値を比較すると、すべての数値で t_0 の方が大きい。したがって、水準 0.1% で母平均値に有意差が認められる。

上記の検定結果から、A薬は旧薬Bに比べ、症状消失に関して優れた効果を持つものと考えてよいだろう。

【解説 1・1】

通常、「2つの母平均値の差の検定」は、2群の母分散が「等しい場合」と「等しくない場合」の2通りに分けて考える必要がある。ここでは、母分散が等しいものと仮定できる場合について、まず説明する。母分散が等しいことを仮定しない場合、もしくは等分散の検定（後述）で有意差が認められた場合については、別に説明する。

2群 A, B の標本数を m, n とし、特定の変数の母集団での分布が正規分布に従い、そこから無作為に抽出されたデータを A 群では x_1, x_2, \dots, x_m とし、B 群では y_1, y_2, \dots, y_n とする。A 群の平均値を \bar{x} とし、不偏分散を s_x^2 とすると、

$$\bar{x} = \frac{1}{m} \sum_{i=1}^m x_i, \quad s_x^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2 \quad (1.1)$$

であり、同様に B 群についても平均値を \bar{y} とし、不偏分散を s_y^2 とすると、

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (1.2)$$

である。

2群での母分散が等しいと仮定できる場合、以下の統計量を求める。すなわち、

$$t_0 = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2} \cdot \left(\frac{1}{m} + \frac{1}{n}\right)}} \quad (1.3)$$

を求める。 t_0 は、自由度 ν 、

$$\nu = m + n - 2 \quad (1.4)$$

の t 分布に従う。したがって、 t 分布表より両側 α 点（通常 5% か 1%）の値 $t_{\alpha/2}(\nu)$ を求め、 t_0 と比較する。

$$t_0 \geq t_{\alpha/2}(\nu)$$

ならば、水準 α で有意差があるものとし、そうでなければ、2群の母平均値に差があるとはいえないくなる。

【検定法 1・2】

例題 1・1 について、ここでは2群の母分散が等しいことを、あらかじめ仮定せずに、母平均値の差の検定を行う。

まず、2群の母分散が等しいか否かについて、「等分散の検定」を行う。2つの分散の比は(1.5)式より、

$$F_0 = \frac{1 \cdot 2^2}{0.8^2} = 2.25$$

となる。 F_0 は自由度 (49, 49) の F 分布に従うので、F 分布の上側 2.5% 点を数表より求めると、約 1.76 である。 F_0 の値はこれより大きいので、両側 5% 水準で等分散の仮説は棄却される。すなわち、2群の母分散には有意差がある。この結果は、A薬とB薬の効果（症状消失までの日数）の分布のばらつき具合が異なることを示している。したがって、この結果のみからでも2薬剤の効果は異なる結論できるが、分布の代表値である平均値についての検定は以下のようになる。

上記の結果から、2群の母分散を等しいものと仮定することはできなくなつたので、前述のように(1.3)式を用いるのは、もはや適切ではない。ここでは、ウエルチの検定を行う。(1.6)式より、

$$t_0 = \frac{|3.1 - 5.3|}{\sqrt{\frac{1 \cdot 2^2}{50} + \frac{0.8^2}{50}}} \doteq 10.786$$

を求める。

t_0 の自由度 ν は、(1.7)式より、

$$\nu = \frac{\left(\frac{1 \cdot 2^2}{50} + \frac{0.8^2}{50}\right)^2}{\frac{1 \cdot 2^4}{50^2 \times (50-1)} + \frac{0.8^4}{50^2 \times (50-1)}} \doteq 85.4$$

となる。自由度 85 の t 分布表の両側 0.1% 点を数表から求めると、約 3.411 である。 t_0 はこれより大きいので、等分散を仮定した場合と同様に、高度に有意差が認められる。

上記の結果から、A薬はB薬に比べ、平均としての効果は優れているが、効果の現われ方について、より個体差が大きいものと考えられる。

【解説 1・2】

前述の解説 1・1 と同様に、2群 A, B について標本数を m, n とし、データの表記なども同一とする。A 群では、変数の分布が母平均 μ_A 、母分散

σ^2_A の正規分布に従い、B群では母平均 μ_B 、母分散 σ^2_B の正規分布に従うものとする。

解説 1・1 で述べた方法は、2つの母分散が「 $\sigma^2_A = \sigma^2_B$ 」という場合に、仮説「 $\mu_A = \mu_B$ 」を検定する方法であつた。したがつて、この方法を用いるためには、あらかじめ母分散が等しいことを、確認しておく必要がある。特定の理論により、等分散が仮定できる場合以外は、以下に述べる「等分散の検定」がよく用いられる。

仮説「 $\sigma^2_A = \sigma^2_B$ 」を検定するためには、

$$F_0 = \frac{s^2_x}{s^2_y} \quad (1.5)$$

を求める。ただし、 s^2_x 、 s^2_y は (1.1)、(1.2) 式で定義される各群の不偏分散である。

F_0 は自由度 $(m-1, n-1)$ の F 分布に、また $1/F_0$ は自由度 $(n-1, m-1)$ の F 分布にそれぞれ従う。したがつて、両側 α の確率で検定を行う場合、自由度 (ν_1, ν_2) の F 分布の上側 $\alpha/2$ 点を $F_{\alpha/2}(\nu_1, \nu_2)$ のように表すと、

$$F_0 > F_{\alpha/2}(m-1, n-1)$$

$$\text{または、} \quad 1/F_0 > F_{\alpha/2}(n-1, m-1)$$

ならば仮説を棄却すればよい。そうでなければ、仮説を採択し、等分散を仮定して、解説 1・1 で述べた方法により、仮説「 $\mu_A = \mu_B$ 」を検定すればよい。ただし、(1.5) 式により F_0 を計算する場合、 $F_0 > 1$ になるように、分子と分母を設定すればよい。

ところで、上記の手順で検定を二度行くと、各検定の有意水準を α とすると、全体の有意水準は 2α となる。したがつて、全体の有意水準を α とするためには、「等分散の検定」も「母平均値の差の検定」も各水準を $\alpha/2$ とする必要がある。すなわち、仮に全体の有意水準を 5% にしたい場合、 F_0 の比較には F 分布の上側 1.25% 点を用い、 t_0 の比較には t 分布の両側 2.5% 点を用いることになる。なお、このように有意水準を設定することは、等分散が仮定されやすくなり、また母平均値の差が有意になりにくくなることを意味している。

2つの母分散について等分散性を仮定せずに、仮説「 $\mu_A = \mu_B$ 」を検定する問題は、ペーレンス・フィッシャー問題として、歴史的に有名なもので

あり、興味深いものであるが、ここでは一般によく用いられている「ウェルチの方法」を紹介する。

等分散の前提なしに、仮説「 $\mu_A = \mu_B$ 」を検定するためには、

$$t_0 = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{s^2_x}{m} + \frac{s^2_y}{n}}} \quad (1.6)$$

を求める。 t_0 は t 分布に従うが、その自由度 ν は、

$$\nu = \frac{\left(\frac{s^2_x}{m} + \frac{s^2_y}{n}\right)^2}{\frac{s^4_x}{m^2(m-1)} + \frac{s^4_y}{n^2(n-1)}} \quad (1.7)$$

となる。ここで、自由度 ν は検定法 1・2 で示したように、整数になるとは限らない。 ν の値が 10 未満の場合、特別な表として「小数自由度の t 分布表」が用意されている。また、 ν が 10 以上ならば、小数部を切り捨て、整数部のみを自由度として採用すればよい。検定の方法は、等分散が仮定できる場合とまったく同一に行えばよい。なお、標本数 m, n が等しい場合、(1.3) 式と (1.6) 式による t_0 の値は、完全に一致する。

2. Wilcoxon の順位和検定

例題 1・3

急性腎盂腎炎患者に対して、無作為に A 薬と B 薬を投与し、下熱日数を調べたところ、第 1 表の「日数欄」に示したようなデータが得られた。このデータから、A 薬と B 薬では下熱日数に差があるといえるだろうか。

【検定法 1・3】

第 1 表のデータで明らかなように、観察期間内で下熱しない例がある。このため、測定値をそのまま用いて、母平均値の差の検定を行うことはできない。ここでは、測定値に順位を付け、それをもとに検定を行う。

第 1 表に示したように、A 薬と B 薬の区別なく、全データについて小さい順に順位を付ける。ただし、同順位のものについては、第 1 表のように、その中間の順位を付けるものとする。すなわち、下熱日数が 2 日というデータが、A 薬で 3 つ、B 薬で 1 つあり、これらについては、1, 2,

第1表 A薬とB薬の投与による下熱日数

A 薬		B 薬	
日 数	順 位	日 数	順 位
2	2.5	2	2.5
2	2.5	3	7
2	2.5	3	7
3	7	3	7
3	7	4	12
4	12	4	12
4	12	5	15
4	12	8	17.5
6	16	10	19
8	17.5	*	22
*	22	*	22
*	22	*	22
計	135	計	165

(*は下熱せず)

3, 4 の順位がつくが、その中間の順位である2.5を4つのデータに与えることにした。他の場合も同様である。なお、このように同順位がある場合を、「タイ」のある場合という。

A薬について、(1.8)式で定義される順位和 R_A^* は第1表に示したように、

$$R_A^* = 2.5 + 2.5 + 2.5 + 7 + \dots + 22 + 22 = 135$$

となり、同様にB薬の順位和 R_B^* は165となる。タイのない場合、標本数が小さければ、上記の順位和の小さい方と、「Wilcoxon の順位和検定用の数表」の両側5%か1%について、対応する標本数についての値と比較し、求めた順位和が大きくなければ、有意差ありとすればよい(数値表によつては、大きい方の順位和を用いて、この逆に検定を行う)。

場合によつては、(1.12)式より、

$$U_A^* = 135 - 12 \times 13 / 2 = 57$$

$$U_B^* = 165 - 12 \times 13 / 2 = 87$$

のように、Mann-Whitney のU統計量を求め、Wilcoxonの順位和検定と同様に、数値表と比較して検定を行うこともできる。しかし、ここではデータにタイがあるので、正規近似により有意差検定を行う。

A薬、B薬群ともに標本数は12であるから、順位和 R_A^* の期待値 $E(R_A^*)$ は(1.19)式より、

$$E(R_A^*) = 12 \times (24 + 1) / 2 = 150$$

となり、その分散 $V(R_A^*)$ は(1.20)式より、

$$V(R_A^*) = \frac{12 \times 12}{12 \times 24 \times 23} \times [(24^3 - 24) - \{(4^3 - 4) + (5^3 - 5) + (5^3 - 5) + (2^3 - 2) + (5^3 - 5)\}] \div 290.739$$

となる。これを、(1.22)式に代入することで、

$$z_0 = \frac{|135 - 150|}{\sqrt{290.739}} \div 0.880$$

と求められる。標準正規分布の両側5%点は1.96であるから、 z_0 はこれより小さく、有意差は認められない。

上記の結果から、A薬とB薬の投与による下熱日数の分布に差はなく、したがつてその効果に違いがあるとは考えられないことになる。

【解説 1・3】

A群とB群の無作為抽出標本数を m, n とし、 $N = m + n$ とする。特定変数に関するA群のデータを x_1, x_2, \dots, x_m とし、B群では y_1, y_2, \dots, y_n としよう。これらのデータは、血圧、体重などのように定量的なものであつても、初めから順位として得られたものであつてもよい。もしくは、「著効」・「有効」などのようにカテゴリ間に自然な順序のあるものでも構わない。

全体でN個のデータを、大きさの順に並べ、1からNまで順位を付ける。ここでは、まず同順位がない、すなわちタイのない場合について説明する。

A群でのデータの順位を s_1, s_2, \dots, s_m とし、B群での順位を r_1, r_2, \dots, r_n とする。A群、B群での順位和をそれぞれ R_A, R_B とすると、

$$R_A = s_1 + s_2 + \dots + s_m$$

$$R_B = r_1 + r_2 + \dots + r_n \tag{1.8}$$

となる。ここで、順位は1からNまでなので、

$$R_A + R_B = N(N + 1) / 2 \tag{1.9}$$

の関係が成り立つ。ここで、説明を簡単にするため、 $R_A > R_B$ であるものとする。

2群の分布の形状は同一であるが、その位置に差があるかを検定するためには(一方がより高い方に分布がずれているか否か)、適当な定数 c_1, c_2 を用いて、

$$R_A \geq c_1 \text{ または } R_B \leq c_2 \tag{1.10}$$

のとき、有意差ありと検定できる。これが「Wil-

coxon の順位和検定」と呼ばれる方法である。c₁, c₂ は有意水準を α としたとき, (1.10) 式が成り立つ確率が等しいように設定される。すなわち,

$$\Pr(R_A \geq c_1) = \Pr(R_B \leq c_2) = \alpha/2 \quad (1.11)$$

となるように定められる (両側検定)。実際には α とし、5% や 1% などについて、2 群の標本数が 20 以下ぐらいまでの場合、c₁, c₂ をいくつにして検定するかを示す数値表が用意されている。また、用いる数値表によつては、以下のように順位和を変換して用いることもある。

A 群の標本数は m なので、順位和 R_A から期待される最小値を引くと、

$$U_A = R_A - m(m+1)/2 \quad (1.12)$$

となる。同様に、B 群については、

$$U_B = R_B - n(n+1)/2$$

となる。U_A, U_B は「Mann-Whitney の U 統計量」と一致する。U 統計量による検定も、適当な定数 u₁, u₂ を用いて、

$$U_A \geq u_1 \text{ または } U_B \leq u_2 \quad (1.13)$$

のとき、有意差ありとすればよい。

実際の検定では、用いる数表が順位和に関するものか U 統計量に関するものかにより、上記のうち一方の方法を選択する必要がある。なお、数表が順位和や U 統計量に関して、小さい方について作成されていることが多いので、応用上注意して欲しい。

標本数が比較的大きい場合、以下のような正規分布による近似を用いて、検定を行うことができる。A 群の順位和 R_A の期待値は、

$$E(R_A) = m(N+1)/2 \quad (1.14)$$

となり、分散は、

$$V(R_A) = mn(N+1)/12 \quad (1.15)$$

となる。B 群については、たんに上式の m と n を交換するだけでよい。また、U 統計量についての期待値は、

$$E(U_A) = mn/2 \quad (1.16)$$

となり、分散は順位和 R_A の分散と一致し、(1.15) 式で与えられる。

期待値と分散を用いて、

$$z_0 = \frac{|R_A - E(R_A)| - 0.5}{\sqrt{V(R_A)}}$$

$$= \frac{|U_A - E(U_A)| - 0.5}{\sqrt{V(U_A)}} \quad (1.17)$$

を求め、標準正規分布の両側 α 点 z_{α/2} と比較し、z₀ が大ならば有意差ありとすればよい。実際には、両側 5% 点 1.96 が 1% 点 2.58 を用いて比較すればよい。なお、(1.17) 式の分子の 0.5 は、離散的な順位データを連続的な正規分布へあてはめる場合、近似がより良くなるように行うもので、「連続性の補正」とよばれるものである。大部分の場合、この補正により近似が良くなるが、有意確率が小さくなると補正をしないほうがよい場合もある。また、2 群の標本数のうち小さい方が 4 程度でも、上記の正規近似はかなりよく、8 以上ならば実用上十分である。

次に、タイのある場合について、若干の説明を加える。一般に、順位和検定もしくは U 統計量による検定のための数値表は、タイのない場合について作成されている。したがって、タイのある場合にはそれを使用することはできない。そのような場合、自分で順位和の分布をすべて数え上げ、確率計算をするか、以下のように標準正規分布による近似を用いることになる。

いま N 個のデータが k 個の異なる値をとるものとする。d₁ 個は最小値に等しく、d₂ 個が次に大きく、以下順に d_k 個が最大値に等しいものとする。i 番目に大きな値をとるデータには、その中間順位として s_i* を与えるものとしよう。すなわち、

$$s_i^* = (1 + d_i)/2 + \sum_{j=1}^{i-1} d_j \quad (i=1, 2, \dots, k) \quad (1.18)$$

である。A 群の順位和を R_A* とすると、その期待値は (1.14) 式と同様に、

$$E(R_A^*) = m(N+1)/2 \quad (1.19)$$

となり、分散は、

$$V(R_A^*) = \frac{mn}{12N(N-1)} \{ (N^3 - N) - \sum_{i=1}^k (d_i^3 - d_i) \} \quad (1.20)$$

となる。タイがない場合、d₁ = d₂ = … = d_N = 1 となり、上式は (1.15) 式と一致する。また、U 統計量については、

$$E(U_A^*) = mn/2, \quad V(U_A^*) = V(R_A^*) \quad (1.21)$$

の関係が成り立つ。検定のためには、

$$z_0 = \frac{|R_A^* - E(R_A^*)|}{\sqrt{V(R_A^*)}} = \frac{|U_A - E(U_A^*)|}{\sqrt{V(U_A^*)}} \quad (1.22)$$

を求め、タイのない場合と同様の手順で検定すればよい。なお、中間順位を許しているため、連続性の補正は行わない。

上式による正規近似は、比 $d_1/N, \dots, d_k/N$ が 1 に近づくとも精度が悪くなる。したがって、特定のタイのデータの個数が極端に多い場合、正規分布による近似を用いるべきではない。

例題 1・4

第2表は、2種の薬剤A、Bを慢性前立腺炎患者に無作為に投与し、その効果を調べたものである。このデータから、AとBの薬効について有意差があるか検定せよ。

第2表 A薬とB薬の症状改善について

	著効	有効	不変	悪化	計
A薬	8	10	6	1	25
B薬	3	9	10	4	26
計	11	19	16	5	51

【検定法 1・4】

第2表のようなクロス表は、一般に χ^2 検定がよく用いられる。しかし、効果を表わす4つのカテゴリに順序が付けられると考えられるので、タイのある場合の Wilcoxon の順位と検定を行うことにする。

「著効」には (1.18) 式を用いて、順位として 6 を、同様に「有効」には 21、「不変」には 38.5、「悪化」には 49 を与える。A薬群の順位和は、

$$R_A^* = 8 \times 6 + 10 \times 21 + 6 \times 38.5 + 1 \times 49 = 538$$

となり、その期待値は (1.19) 式より、

$$E(R_A^*) = 25 \times (51 + 1) / 2 = 650$$

となる。また、分散は (1.20) 式より、

$$\begin{aligned} V(R_A^*) &= \frac{25 \times 26}{12 \times 51 \times 50} \times [(51^3 - 51) - \{(11^3 - 11) \\ &\quad + (19^3 - 19) + (16^3 - 16) + (5^3 - 5)\}] \\ &\doteq 2554.118 \end{aligned}$$

となる。これを (1.22) 式に代入すると、

$$z_0 = \frac{|538 - 650|}{\sqrt{2554.118}} \doteq 2.216$$

と求められる。標準正規分布の両側 5% 点は 1.96 であり、 z_0 はこれより大きく、したがって有意差が認められる。この結果から、A薬はB薬に比べ、症状改善に有効なものと考えられる。なお、後述するクロス表の検定として χ^2 検定を用いると、 χ^2 値は 5.108 となり、その有意確率は 0.164 (16.4%) となり、有意とはならない。Wilcoxon の順位と検定では、カテゴリ間の順序に関する情報を用いるため、標本数が小さい場合に、有効な方法である。しかし、処理効果について順序を付けず、たんに2群の効果の差を調べたいのなら、通常の χ^2 検定を用いた方がよい。

最後に、実際に応用する場合に必要な統計数値表として文献の1を、また各手法についてもつと知りたい読者のためには、2~4を紹介する。

文献

- 1) 統計数値表編集委員会編：簡約統計数値表，日本規格協会，1977.
- 2) 豊川裕之，柳井晴夫編著：医学・保健学の例題による統計学，現代数学社，1982.
- 3) E.L. レーマン原著，鍋谷清治他訳：ノンパラメトリックス，森北出版，1978. (E.L. Lehmann: Nonparametrics, Holden-Day, 1975)
- 4) 竹内 啓，大橋靖雄：統計的推測—2標本問題，日本評論社，1981.