１３章 対応のある２群間の量的データの検定

13.1 検定手順

この章では対応がある場合の量的データの検定方法について学びます。この場合も図 13.1 のように最初に正規分布に従うかどうかを調べます。正規性が認められた場合は対応がある場合の t 検定、正規性が認められない場合はウィルコクソン（Wilcoxon）の符号付き順位和検定を行ないます。１１章で述べた検定方法と似ていますが、ここでは対応のあるデータ同士を引き算した値を用いて判断します。正規性はこの差を用いて確認し、正規分布と異なるといえないと判定された場合、一応正規分布とみなしますが、怪しい場合は両方の検定を試すことをお勧めします。

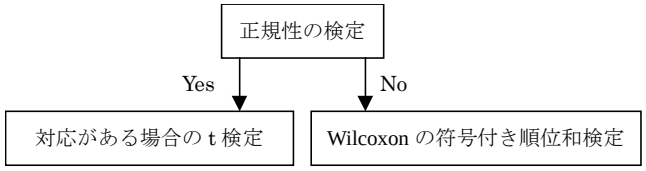

図 13.1 対応のある量的データの検定手法

13.2 2 群間の等平均の検定（対応あり・正規性あり）

ここでは正規性が認められたと仮定して、対応がある場合の平均の比較方法について説明をします。以下の例を見て下さい。

例

規模の等しい８つの支店で、ある商品の陳列位置を変える前と後とで売上高（千円）を比較したところ、以下の結果を得た。各標本が正規分布するとして有意水準 5%で差があるかどうか判定せよ。

前	385	402	320	383	504	417	290	342
後	396	373	431	457	514	405	380	396

理論　対応がある場合の t 検定

正規分布する対応のある標本１と標本２の母平均 μ_1，μ_2 を比較し、差があるかどうか有意水準 $\alpha \times 100\%$ で判定する。

対応する 2 群のデータの差（$z_i = X_{1i} - X_{2i}$）について、データ数 n，平均 $\bar{z}$，不偏分散 u_z^2 とする。

帰無仮説 H_0： $\mu_1 = \mu_2$　　平均に差がない

対立仮説 H_1： $\mu_1 \neq \mu_2$　　平均に差がある（両側検定）

H_0 のもとで　　$t = \dfrac{\sqrt{n}\,\bar{z}}{u_z} \sim t_{n-1}$ 分布　　(13.1)

$p = tdist(|t|, n-1, 2)$ として、 $p < \alpha$ ならば、H_0 を棄却して H_1 を採択する。

解答

まず、標本 1 を変更前、標本 2 を変更後とし、対応するデータ間の差を求めて以下のような表を作ります。

前	385	402	320	383	504	417	290	342
後	396	373	431	457	514	405	380	396
差	-11	29	-111	-74	-10	12	-90	-54

これから、差の平均と標準偏差を求めます。

$$n = 8,\ \bar{z} = -38.625,\ u_z = 50.82726$$

これを用いて検定統計量を計算すると以下のようになります。

$$t = -2.149398$$

自由度は 8-1=7 ですから、検定確率値は以下で与えられます。

$$p = tdist(2.149398, 7, 2) = 0.068675 \cong 0.069$$

$p > 0.05$ より、平均に差があるとはいえないと判定されます。

解説

対応のある場合とない場合で検定の方法が違いますが、対応のない場合の手法は単に 2 群に分かれているだけですから、対応があってもなくても利用可能のはずです。ではなぜ特別な方法を利用するのでしょうか。対応がある場合は、1 対のデータに対して、一方の群でその大きさが大きければ、他方の群でも大きい場合が多いように思われます。例えば、キャンペーン前後の売上高の比較をする場合、元々売上の大きい店舗はキャンペーン後も売上が大きい可能性が高いと思ってもらえれば良いでしょう。この性質が 2 群の対応するデータの差を取るという検定方法を利用する理由です。即ち、対応するデータの差を取ると、上の性質からデータの分散がより小さくなり、差の判別が容易になると思えるからです。詳細は次の数学的解説に譲りましょう。

数学的解説　[Skip OK]

まず検定統計量 t が自由度 $n-1$ の t 分布になることを説明しておきましょう。2 群の

データ X_{1i}, X_{2i} はそれぞれ独立に $N(\mu,\sigma^2)$ 分布に従うとします。その差を取ると、$z_i \sim N(0, 2\sigma^2)$ 分布となります。またこの平均 $\bar{z}$ をとって、$\bar{z} \sim N(0, 2\sigma^2/n)$ 分布です。これから、以下の関係となります。

$$\sqrt{n}\bar{z}/\sqrt{2}\sigma \sim N(0,1) \text{ 分布}$$

次に z_i の分散 u_z^2 を取ると以下の性質があります。

$$\chi^2 = \frac{(n-1)u_z^2}{2\sigma^2} \sim \chi_{n-1}^2 \text{ 分布}$$

これらを合せて以下のようになります。

$$t = \frac{\sqrt{n}\bar{z}}{u_z} = \frac{\sqrt{n}\bar{z}}{\sqrt{2\sigma^2\chi^2/(n-1)}} = \frac{\sqrt{n}\bar{z}/\sqrt{2}\sigma}{\sqrt{\chi^2/(n-1)}} \sim t_{n-1} \text{ 分布}$$

この計算には 2 群の分散が等しいという暗黙の了解があります。実際に対応のあるデータでは分散に差はないと思われますが、やはり注意が必要です。

さて、解説のところでも述べたように、ここの検定方法を対応がない 2 群の平均の検定と比べてみましょう。対応がない場合の検定統計量は以下で与えられます。

$$t' = \sqrt{\frac{n_1 n_2}{n_1+n_2}} \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{(n_1-1)u_1^2 + (n_2-1)u_2^2}{n_1+n_2-2}}} \sim t_{n_1+n_2-2} \text{ 分布}$$

対応のある場合、データ数は $n_1 = n_2 = n$ ですから、この検定統計量は以下のようになります。

$$t' = \frac{\sqrt{n}\bar{z}}{\sqrt{u_1^2+u_2^2}} \sim t_{2(n-1)} \text{ 分布} \qquad \text{但し、} \bar{z} = \frac{1}{n}\sum_{i=1}^{n}(X_{1i} - X_{2i}) = \bar{x}_1 - \bar{x}_2$$

この分母の 2 乗と u_z^2 とを比較してみましょう。

$$\begin{aligned} u_z^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 = \frac{1}{n-1}\sum_{i=1}^{n}[(X_{1i} - X_{2i}) - (\bar{x}_1 - \bar{x}_2)]^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}[(X_{1i} - \bar{x}_1)^2 + (X_{2i} - \bar{x}_2)^2 - 2(X_{1i} - \bar{x}_1)(X_{2i} - \bar{x}_2)] \\ &= u_1^2 + u_2^2 - 2u_{12} \end{aligned}$$

ここに、u_{12} は X_{1i} と X_{2i} の共分散と呼ばれます。一般にこの共分散は正負どちらも可能性がありますが、対応のある場合には、大きいデータ同士、小さいデータ同士が対になることが多いので、正の値になることが多いと思われます。その場合には、以下の関係が導かれます。

$$u_z^2 = u_1^2 + u_2^2 - 2u_{12} < u_1^2 + u_2^2 \quad \rightarrow \quad t > t'$$

このことから、検定統計量は 2 群のデータの差を取った t の方が大きくなると思われます。即ち、検定確率も t を用いた方が小さく、平均の差は出易いでしょう。但し、自由度は統計量 t が t' の半分になりますので、差は出にくくなります。しかし、自由度が大きい場合、この差は t の値の差に比べて影響が少ないですから、結局統計量 t を用いる方が有利という結論になります。もちろんデータによってはこのようにならない場合もあるでしょう。

13.3 2 群間の中央値の検定（対応あり・正規性なし）

対応があり、正規分布に従わない場合は、11.4 節で学んだウィルコクソン（Wilcoxon）の符号付き順位和検定をデータの形を変えて適用します。即ち、以前はデータから母集団の中央値を引いて新たな変数を作りましたが、今度は前節で述べたように、対応するデータ同士の差を新たな変数とします。それでは例を見ていきましょう。

例

ある商品の陳列位置を変える前と後とで売上高（千円）を規模の等しい８つの支店で比較したところ、以下の結果を得た。各標本が正規分布しないものとして有意水準 5%で売上高に差があるかどうか判定せよ。

前	385	402	320	383	504	417	290	342
後	396	373	431	457	514	405	380	396

理論　（対応がある場合の）Wilcoxon の符号付き順位和検定

任意の分布に従う対応のある標本１と標本２の母集団の中央値 m_1， m_2 を比較し、差があるかどうか有意水準 $\alpha \times 100\%$ で判定する。

帰無仮説 H_0： $m_1 = m_2$　　中央値に差がない

対立仮説 H_1： $m_1 \neq m_2$　　中央値に差がある（両側検定）

対応する各標本の差（$z_i = X_{1i} - X_{2i}$）について、0 を除いて $|z_i|$ の小さい順に順位 r_i を付け、z_i の正負で 2 群に分ける。但し、同数値の場合は、順位平均を取る。例えば、5 位が 2 つの場合は、両方 (5+6)/2=5.5 とする。

各群のデータ数を r, s（$n = r + s$）、順位和を R_r， R_s とし、順位和の小さい方を R とする。

データ数が 50 以下のとき

補遺３の数表を参照し、両側確率をαとして$R \le R_1$のとき、中央値に差があると判定する。

データ数が 50 より多いとき

H_0のもとで　$z = \dfrac{|R - n(n+1)/4| - 1/2}{\sqrt{n(n+1)(2n+1)/24}}$　～　$N(0,1)$分布（正の部分）

(13.2)

$p = 2 \cdot (1 - normsdist(z))$として、$p < \alpha$のとき、$H_0$を棄却して$H_1$を採択する。

解答

データをもとに以下のような表を作ります。11.4 節とは 2 群の差を取るまでは違いますが、差の絶対値の大きさで順位を付ける方法は同じです。訂正順位については何度も出ましたが、同順位の場合は平均を取ります。

前	後	差	｜差｜	訂正順位
385	396	-11	11	2
402	373	29	29	4
320	431	-111	111	8
383	457	-74	74	6
504	514	-10	10	1
417	405	12	12	3
290	380	-90	90	7
342	396	-54	54	5

これより、小さい方の順位和は、$R = 7$

補遺３数表より、$R = 7 > R_1 = 3$　であり、中央値に差があるとはいえないと判定します。

１４章 相関係数の検定

14.1 ピアソンの相関係数

2 変量の関係を表わす統計量として、5.4 節で相関係数の話をしましたが、ここでは相関の有無を検定するということに主眼を置いて話を進めます。

以前述べた相関係数はピアソン(Pearson)の相関係数と呼ばれ、変数 x と y がどの程度 $y = ax + b$ の関係で表わされるのかを知るための指標でした。直線状に乗れば相関係数は ± 1 に近く、そうでなければ 0 に近くなりました。しかし、相関があると判断するためには相関係数の値はどの程度必要なのでしょうか。また、標本データからの偶然性はどのように評価するのでしょうか。標本データから相関係数を求めると通常 0 でない値が出ます。しかしこれは偶然でないと言えるのでしょうか。ここではこの評価の方法を説明します。例を見ていきましょう。

例

２つの商品 A, B の地域別使用率 (%) のデータは下の表の通りである。それぞれの商品の使用率に相関が認められるか。正規性を仮定して、有意水準 5%で判定せよ。

A(%)	33	24	30	50	42	15	15	56	13	45	44	21	18	31	27	40
B(%)	20	34	50	20	58	23	12	34	26	56	42	5	25	51	19	27

理論

2 変数が 2 変量正規分布に従うとき、母相関係数 ρ が 0 と異なるか（差があるか）を判定する。データ数を n, 相関係数を r とする。

帰無仮説 H_0： $\rho = 0$　　相関係数は 0 である

対立仮説 H_1： $\rho \neq 0$　　相関係数は 0 でない

H_0 のもとで　　$t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$ 分布　　(14.1)

$p = tdist(|t|, n-2, 2)$ として、 $p < \alpha$ ならば、H_0 を棄却して H_1 を採択する。

解答

相関係数 r は、Excel では correl(範囲 1,範囲 2) 関数で与えられます。これから次のような値を得ます。

$$r = 0.453786, \quad n = 16$$

これを用いて検定統計量を計算すると以下のようになります。

$$t = 1.905387$$

自由度を 16-2=14 として、検定確率は以下となります。

$$p = tdist(1.905387, 14, 2) = 0.077476 \cong 0.077$$

$p > 0.05$ より、相関係数は 0 と異なるといえないと判定されます。

解説

実はこの検定には落とし穴があります。相関係数が 0 でないことを示すだけですから、データ数を多く取るとほんの少し直線に近付くだけで、相関があると判定されてしまいます。実際に相関があるというのは、検定で 0 でないと結論が出た上で、どの程度の数値になるべきかを分析する本人が判断することだと思います。これに関連して、相関係数の検定には、相関係数が 0 以外のある値と異なるかどうかを判定するものもありますが、ここでは 0 かどうかに限定して話をしています。

ここで述べるピアソンの相関係数の検定では、データの分布が多変量正規分布に従うことが前提条件です。ただこれを調べるのは難しいと思われますので、それぞれの変量が正規分布することだけでも調べておく必要があります。正規分布しない場合は、次節で述べる順位相関という考え方を用います。

以前にも相関係数の値は 0.5 位欲しいと述べましたが、ここの結果はこれに近い値を出していながら、統計的に 0 でないといえないと結論されています。このように標本によるばらつきというものは相当大きいと思われますので、データ数が多くない場合は十分注意が必要です。検定統計量が自由度 $n-2$ の t 分布に従うことの証明はここでは省略します。

問題

以下の 2 変数間の相関係数を求め、正規性を仮定して、相関係数が 0 と異なるかどうか判定せよ。

変数 1	65	86	78	83	85	89	83	80	85	93	75	85	79	80
変数 2	162	210	224	179	217	230	223	204	224	197	186	189	172	185

解答

$$r = 0.557714 \qquad t = 2.327588 \qquad p = 0.038237 \cong 0.038$$

$p < 0.05$ より、相関係数が 0 と異なるといえる。

14.2 スピアマンの順位相関係数

データに正規性が見られないとき、相関の有無は順位相関係数を用いて検定を行ないます。スピアマン（Spearman）の順位相関係数は、2 群それぞれのデータの小さい順に順位を付け、その順位を用いてピアソンの相関係数を計算したものです。検定はこの相関係数を用いて、前節と同じ統計量で行ないます。例を見ていきましょう。

例

前節の問題で、それぞれの商品の使用率に相関が認められるか。正規性を仮定せずに、有意水準 5%で検定せよ。

理論

一般の分布に従う 2 変数について、順位相関係数 r_s を求め、母集団の順位相関係数 ρ_s が 0 と異なるかどうかを判定する。

データ数 n, 相関係数 r とする。

帰無仮説 H_0： $\rho_s = 0$　　母順位相関係数は 0 である

対立仮説 H_1： $\rho_s \neq 0$ （両側検定）　母順位相関係数は 0 でない

H_0 のもとで　　$t = \dfrac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \sim t_{n-2}$ 分布　　(14.2)

$p = tdist(|t|, n-2, 2)$ として、 $p < \alpha$ ならば、H_0 を棄却して H_1 を採択する。

解答

まず各変量ごとに小さい順に順位を付けます。但し、同順位の場合は以前にも述べたように異なる順位とした場合の平均とします。順位は rank(数値,範囲,順序) 関数か、［データ－並べ替え］メニューを用いると簡単に与えることができます。

A (%)	B (%)	順位 A	順位 B	訂正 A	訂正 B
33	20	10	4	10	4.5
24	34	6	10	6	10.5
30	50	8	13	8	13
50	20	15	4	15	4.5
42	58	12	16	12	16
15	23	2	6	2.5	6
15	12	2	2	2.5	2
56	34	16	10	16	10.5
13	26	1	8	1	8
45	56	14	15	14	15
44	42	13	12	13	12
21	5	5	1	5	1
18	25	4	7	4	7
31	51	9	14	9	14

27	19	7	3	7	3
40	27	11	9	11	9

次にそれらの順位について、correl(範囲 1, 範囲 2) 関数を用いて順位相関係数を計算します。

$$r_s = 0.461312, \quad n = 16$$

これを利用して検定統計量を求めます。

$$t = 1.945443$$

自由度は 16-2=14 ですので、検定確率値は以下のようになります。

$$p = tdist(1.945443, 14, 2) = 0.072084 \cong 0.072$$

$p > 0.05$ より、順位相関係数は 0 と異なるとはいえないと判定されます。

解説

相関を表わす有名な指標には、ピアソンの相関係数、スピアマンの順位相関係数の他に、ケンドール（Kendall）の τ （タウ）と呼ばれるものがあります。これは n 個の全データから $\{x_i, y_i\}$ と $\{x_j, y_j\}$ のように 2 組のデータの組合わせを取り（組合せの数は ${}_nC_2 = n(n-1)/2$）、$x_i > x_j, y_i > y_j$ または $x_i < x_j, y_i < y_j$ のように不等号が同じ向きになっているものの数を P、どちらかの不等号が逆向きになっているものの数を Q として以下のように与えられます。

$$\tau = \frac{P - Q}{n(n-1)/2}$$

ここに $x_i = x_j$ または $y_i = y_j$ となる組の数は P と Q に含めません。この例の場合、$P = 78$, $Q = 39$ で $\tau = 0.325$ となります。一般に τ の値は、$-1 \le \tau \le 1$ となっていますが、これ以上の詳細は省略します。