

12章 対応のない2群間の量的データの検定

12.1 検定手順

この章ではデータ間に1対1の対応のない2つの標本から推定される母集団間の平均値や中央値の比較を行ないます。

検定手法は、図 12.1 のようにまず正規分布に従うかどうかを調べます。但し、この場合は2つの群が共に正規分布に従うことを調べる必要があります。次に2群とも正規分布ならばF検定を用いて等分散であるかどうかを調べます。等分散の場合はt検定、等分散でない場合はウェルチ (Welch) のt検定を行ないます。正規分布でなければノンパラメトリック検定の1つウィルコクソン (Wilcoxon) の順位和検定を行ないます。正規分布に従うかどうかの判定方法は前章で学びましたので、ここでは等分散の検定から話を始めます。

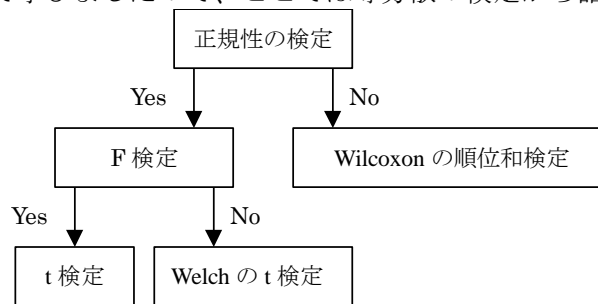


図 12.1 対応のない2群間の量的データの検定手法

12.2 2群間の等分散の検定 (対応なし・正規性あり)

ここではデータの正規性を認めて等分散性の検定について説明します。まず2つの群の不偏分散が求まっているものとして例を見ていきましょう。

例

A機を導入した会社18社(1群)とB機を導入した会社9社(2群)について、機械10台当たりの1年間の故障発生件数を調べ、不偏分散を求めたら以下の結果を得た。

1群 10.68 2群 2.17

分布は正規分布であると仮定して、分散に差があるといえるか有意水準5%で判定せよ。

理論 F検定

正規分布する母集団から抽出した標本1と標本2について、それぞれの母分散 σ_1^2, σ_2^2 を比較し、差があるかどうか有意水準 $\alpha \times 100\%$ で検定を行う。

データ数 n_1, n_2 , 不偏分散 u_1^2, u_2^2 (但し、 $u_1^2 > u_2^2$ とする) とする。

帰無仮説 $H_0: \sigma_1^2 = \sigma_2^2$ 分散に差がない

対立仮説 $H_1: \sigma_1^2 > \sigma_2^2$ 分散に差がある

$$H_0 \text{のもとで} \quad F = \frac{u_1^2}{u_2^2} \sim F_{n_1-1, n_2-1} \text{分布} \quad (12.1)$$

$p = fdist(F, n_1 - 1, n_2 - 1)$ として、 $p < \alpha$ ならば H_0 を棄却して H_1 を採択する。

注) $p = fdist(x, d_1, d_2)$ 片側検定

解答

$u_1^2 > u_2^2$ とするため、 $n_1 = 18$, $n_2 = 9$, $u_1^2 = 10.68$, $u_2^2 = 2.17$ のように選びます。これから検定統計量 F を計算します。

$$F = \frac{10.68}{2.17} = 4.921659$$

自由度は、 $18-1=17$ と $9-1=8$ で、検定確率値は以下となります。

$$p = fdist(4.921659, 17, 8) = 0.013776 \cong 0.014$$

$p < 0.05$ より、分散に差があるといえると判定されます。

解説

この検定は計算が非常に簡単です。但し、分散の比 u_1^2/u_2^2 が 1 より大きくなるように標本 1 と標本 2 を選ぶ必要がありますので注意して下さい。その際自由度を求めるデータ数も間違えないようにして下さい。

検定結果として分散に差があると判定された場合は問題ありませんが、差があるといえないと判定された場合、正確には「このデータでは差があるといえない」ですから、積極的に等分散性を示しているわけではありません。しかし、一般には不本意ながらこれで等分散性が示されたことにしています。検定確率値が有意水準に近い場合は、 t 検定と Welch の t 検定と両方試してみることをお勧めします。

数学的解説

この理論を簡単に調べてみましょう。各変量 X_{1i} ($i=1, \dots, n_1$) , X_{2j} ($j=1, \dots, n_2$) について、それぞれ独立で以下の正規分布に従うとします。

$$X_{1i} \sim N(\mu_1, \sigma^2) \text{ 分布}, \quad X_{2j} \sim N(\mu_2, \sigma^2) \text{ 分布}$$

これから、

$$\chi_1^2 = \frac{(n_1 - 1)u_1^2}{\sigma^2} = \sum_{i=1}^{n_1} \frac{(X_{1i} - \bar{x}_1)^2}{\sigma^2} \sim \chi_{n_1-1}^2 \text{ 分布}$$

$$\chi_2^2 = \frac{(n_2 - 1)u_2^2}{\sigma^2} = \sum_{j=1}^{n_2} \frac{(X_{2j} - \bar{x}_2)^2}{\sigma^2} \sim \chi_{n_2-1}^2 \text{ 分布}$$

となります。ここに、 \bar{x}_α ($\alpha=1, 2$) は標本平均です。自由度で割ったこれら 2 つの変量の比をとると、以下のようになります。

$$F = \frac{u_1^2}{u_2^2} = \frac{\chi_1^2 / (n_1 - 1)}{\chi_2^2 / (n_2 - 1)} \sim F_{n_1-1, n_2-1} \text{ 分布}$$

問題

以下の標本データの母分散には差があるといえるか。有意水準 5% で検定せよ。

標本 1 112, 106, 101, 112, 102, 98, 108, 95, 101, 90, 110, 97, 95, 105, 101, 113, 114, 91

標本 2 98, 88, 105, 99, 96, 93, 109, 106, 103, 87, 107, 102, 97, 91

解答

	データ数	不偏分散
標本 1	18	57.91176
標本 2	14	50.09341

$$F = \frac{57.91176}{50.09341} = 1.156076$$

$$p = \text{fdist}(1.156, 17, 13) = 0.401544 \cong 0.402$$

$p > 0.05$ より、分散に差があるといえない。

12.3 2 群間の等平均の検定（対応なし・正規性あり・等分散性あり）

2 群のデータに対応がなく、正規性が認められ、等分散性が示された場合、等平均の t 検定を行ないます。ここではその方法を学びましょう。

例

ある地域の同性・同年齢の児童について、ある要因の有無による 2 つの集団の体重を調べたところ以下のデータを得た。2 つの集団の平均値に差はあるといえるか。正規性、等分散性を仮定して、有意水準 5% で判定せよ。

	データ数	平均	不偏分散
要因なし	20	40.2	25.5
要因あり	20	36.4	16.0

理論 (student の) t 検定

正規分布する等分散の標本 1 と標本 2 について、それぞれの母平均 μ_1, μ_2 を比較し、差があるかどうか有意水準 $\alpha \times 100\%$ で検定を行う。

データ数 n_1, n_2 , 標本平均 \bar{x}_1, \bar{x}_2 , 不偏分散 u_1^2, u_2^2 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)

$$H_0 \text{ のもとで } t = \frac{\sqrt{\frac{n_1 n_2}{n_1 + n_2}} (\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(n_1 - 1)u_1^2 + (n_2 - 1)u_2^2}{n_1 + n_2 - 2}}} \sim t_{n_1 + n_2 - 2} \text{ 分布} \quad (12.2)$$

$p = \text{tdist}(|t|, n_1 + n_2 - 1, 2)$ として、 $p < \alpha$ ならば H_0 を棄却して H_1 を採択する。

注) $p/2 = \text{tdist}(t, d, 1)$ 検定値 \rightarrow 片側確率

$p = \text{tdist}(t, d, 2)$ 検定値 \rightarrow 両側確率

$t = \text{tinv}(p, d)$ 両側確率 \rightarrow 検定値

解答

$n_1 = 20, n_2 = 20, \bar{x}_1 = 40.2, \bar{x}_2 = 36.4, u_1^2 = 25.5, u_2^2 = 16.0$ として、 t 統計量を求めます。

$$t = \sqrt{\frac{20 \cdot 20}{40}} \frac{40.2 - 36.4}{\sqrt{\frac{19 \cdot 25.5 + 19 \cdot 16.0}{38}}} = 2.637999$$

自由度は $20+20-2=38$ として、検定確率値は以下のようになります。

$$p = \text{tdist}(2.637999, 38, 2) = 0.01202 \cong 0.012$$

$p < 0.05$ ですから、平均に差があるといえると判定します。

数学的解説

ここでは検定統計量 t が自由度 $n_1 + n_2 - 2$ の t 分布に従うことを調べてみましょう。まず、データが群 α 及び個体 i に独立に $X_{ai} \sim N(\mu, \sigma^2)$ 分布であるとします。そのとき、平均値は以下の分布に従います。

$$\bar{x}_1 \sim N(\mu, \sigma^2/n_1) \text{ 分布}, \quad \bar{x}_2 \sim N(\mu, \sigma^2/n_2) \text{ 分布}$$

これから、平均を引き標準偏差で割ると、

$$\frac{\bar{x}_1 - \mu}{\sigma} \sim N(0, 1/n_1) \text{ 分布}, \quad \frac{\bar{x}_2 - \mu}{\sigma} \sim N(0, 1/n_2) \text{ 分布}$$

となり、これらを引いて次のような分布を作ります。

$$\frac{\bar{x}_1 - \bar{x}_2}{\sigma} = \frac{\bar{x}_1 - \mu}{\sigma} - \frac{\bar{x}_2 - \mu}{\sigma} \sim N\left(0, \frac{n_1 + n_2}{n_1 n_2}\right) \text{ 分布}$$

正規化して、まず以下の統計量を与えます。

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\bar{x}_1 - \bar{x}_2}{\sigma} \sim N(0, 1) \text{ 分布}$$

一方、分散には次の関係があることと、

$$\frac{(n_1 - 1)u_1^2}{\sigma^2} \sim \chi_{n_1 - 1}^2 \text{ 分布}, \quad \frac{(n_2 - 1)u_2^2}{\sigma^2} \sim \chi_{n_2 - 1}^2 \text{ 分布}$$

χ^2 分布の性質を用いて、以下のような統計量を作ります。

$$\frac{(n_1 - 1)u_1^2 + (n_2 - 1)u_2^2}{\sigma^2} = \frac{(n_1 - 1)u_1^2}{\sigma^2} + \frac{(n_2 - 1)u_2^2}{\sigma^2} \sim \chi_{n_1 + n_2 - 2}^2 \text{ 分布}$$

これらを用いて t 分布の定義より、検定統計量を得ます。

$$\begin{aligned} & \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)u_1^2 + (n_2 - 1)u_2^2}{n_1 + n_2 - 2}}} \\ &= \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\bar{x}_1 - \bar{x}_2}{\sigma} \bigg/ \sqrt{\frac{(n_1 - 1)u_1^2 + (n_2 - 1)u_2^2}{\sigma^2 (n_1 + n_2 - 2)}} \sim t_{n_1 + n_2 - 2} \text{ 分布} \end{aligned}$$

問題

前節問題の標本データの平均には差があると言えるか。等分散性を認めて、有意水準 5% で判定せよ。

標本 1 112, 106, 101, 112, 102, 98, 108, 95, 101, 90, 110, 97, 95, 105, 101, 113, 114, 91

標本 2 98, 88, 105, 99, 96, 93, 109, 106, 103, 87, 107, 102, 97, 91

解答

	データ数	平均	不偏分散
群 1	18	102.8333	57.91176
群 2	14	98.64286	50.09341

$$t = \frac{\sqrt{18 \cdot 14}}{32} \frac{102.833 - 98.643}{\sqrt{\frac{17 \cdot 57.912 + 13 \cdot 50.093}{30}}} = 1.592559$$

$$p = tdist(1.592559, 30, 2) = 0.121743 \cong 0.122$$

$p > 0.05$ より、平均に差があるといえない。

12.4 2 群間の等平均の検定 (対応なし・正規性あり・等分散性なし)

2 群のデータに対応がなく、正規性が認められ、等分散性が見られない場合は Welch の t 検定と呼ばれる検定方法を利用します。この方法は自由度を求める際に少し面倒な計算が必要です。実際に例を見てみましょう。

例

ほぼ規模の揃った、A機を導入した会社 18 社 (1 群) と B機を導入した会社 9 社 (2 群) について、機械 10 台当りの 1 年間の故障発生件数を調べ、平均と不偏分散を求めたところ以下の結果を得た。正規性があり、異分散であるとして、2 群間の平均に差があるかどうか有意水準 5% で判定せよ。

	平均	不偏分散
1 群	10.56	10.68
2 群	8.22	2.17

理論 ウェルチ (Welch) の t 検定

正規分布する分散の異なる標本 1 と標本 2 について、母平均 μ_1, μ_2 を比較し、差があるかどうか有意水準 $\alpha \times 100\%$ で検定を行う。

データ数 n_1, n_2 , 標本平均 \bar{x}_1, \bar{x}_2 , 不偏分散 u_1^2, u_2^2 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)

$$c = \frac{u_1^2/n_1}{u_1^2/n_1 + u_2^2/n_2} \text{ として、自由度を } d = \frac{1}{\frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1}} \text{ の整数部分とすると}$$

$$H_0 \text{ のもとで } t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{u_1^2/n_1 + u_2^2/n_2}} \sim t_d \text{ 分布}$$

$p = tdist(|t|, d, 2)$ 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、自由度を求めます。

$$c = \frac{10.682/18}{10.682/18 + 2.172/9} = 0.711052, \quad d = 24.88971$$

となるので、自由度を 24 とします。また検定統計量は以下となります。

$$t = \frac{10.56 - 8.22}{\sqrt{10.68/18 + 2.17/9}} = 2.561634$$

これらを用いると検定確率値は以下となります。

$$p = tdist(2.561634, 24, 2) = 0.017123 \cong 0.017$$

$p < 0.05$ より、平均に差があるといえると判定します。

解説

この検定では一般に d の値は整数にはなりません。そこで自由度の値は d の小数点以下を切り捨てにします。例えば上の例で、小数点以下を切り捨てにした場合と切り上げにした場合とで検定確率値の値は以下のように異なります。

$$p = tdist(2.561634, 24, 2) = 0.017123$$

$$p = tdist(2.561634, 25, 2) = 0.016832$$

一般に自由度の大きい場合と小さい場合を比べて、小さい方が検定確率値は大きくなり、差があることを厳しく評価したいなら、小数点以下を切り捨てる方がよいと思われます。

問題

以下の標本データの平均には差があるとみなせるか。正規性・異分散性を認めて、有意水準 5% で判定せよ。

標本 1	358, 469, 397, 350, 329, 446, 393, 379, 443, 348, 455, 332, 311, 424, 420, 354, 353, 390, 434, 430
標本 2	335, 387, 385, 343, 394, 351, 404, 391, 330, 363, 319, 334, 348, 396, 408, 403, 415, 353, 377, 399

解答

	標本 1	標本 2
データ数	20	20
平均	390.75	371.75
不偏分散	2239.461	914.9342

$$c = 0.709949 \quad d = 32.30427 \rightarrow 32$$

$$t = 1.512901$$

$$p = tdist(1.512901, 32, 2) = 0.140119 \cong 0.140$$

$p > 0.05$ より、平均に差があるといえない。

12.5 2 群間の中央値の検定 (対応なし・正規性なし)

ここでは 2 群のデータに正規性が認められない場合について、分布の中央値の差を検定し

ます。中央値も平均値と並び分布の中心を表わす統計量の1つです。この方法はデータを順位データとして扱うことから、差を調べるものは平均ではなく、中央値になりますが、違いをあまり強く意識しないでも良いように思います。また、この検定は分布型によらない検定ですので、正規分布する場合にも適用可能です。但し、差の出易さは前2節で述べたパラメトリック検定に劣ります。それでは実際の例を見てみましょう。

例

あるソフトウェアの販売において、支店の売上伸び率を2つの販売戦略グループで比較したところ、以下の結果が得られた。2群の増加は1群のそれに比べて大きいと言えるか。有意水準5%で判定せよ。

1群 : 6.1, 5.8, 9.2

2群 : 13.5, 16.2, 21.3, 8.1, 16.8

理論 ウィルコクソン(Wilcoxon)の順位和検定

正規分布するとは限らない標本1(データ数 n_1)と標本2(データ数 n_2)について、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 $\alpha \times 100\%$ で検定する。

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)

両群のデータを合せて小さい順に順位を付ける。ただし、同じ値にはそれらが異なると考えた場合の順位の平均値を付ける。但し、 $n_1 \leq n_2$ とする。1群(データ数が少ない方)についてこの順位の合計を取り、 W とする。

$n_2 \leq 20$ の場合

$W \rightarrow$ 確率 $p = \alpha$ (両側検定)として補遺4の数表を参照する。

データ数(n_1, n_2)の組で(U_1/U_2)の値を求め、

$W \leq U_1$ または $W \geq U_2$ であれば、 H_0 を棄却して H_1 を採択する。

$n_2 > 20$ の場合

$$H_0 \text{のもとで } Z = \frac{|W - n_1(n_1 + n_2 + 1)/2| - 1/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}} \sim N(0, 1) \text{ 分布 (正の部分)} \quad (12.4)$$

$p = 2 \cdot (1 - \text{normsdist}(Z)) < \alpha$ であれば、 H_0 を棄却して H_1 を採択する。

解答

まず、データから順位と順位和を求めます。

1群	2群	1群順位	2群順位	群	データ	順位
6.1	13.5	2	5	1	6.1	2
5.8	16.2	1	6	1	5.8	1
9.2	21.3	4	8	1	9.2	4
	8.1		3	2	13.5	5
	16.8		7	2	16.2	6
				2	21.3	8
				2	8.1	3
				2	16.8	7
順位和		7	29			

実際の Excel への入力方法は右表のように全部を1つの列に入力すると順位付けなどが簡単になります。

データ数は 3, 5 ですから、データ数の少ない1群の順位和を求めます。

$$W = 7$$

データ数が 20 以下なので補遺 4 の数表を利用します。数表より、 $n_1 = 3, n_2 = 5$ の場合、両側確率 5% で、 $U_1 = 6, U_2 = 21$ となります。順位合計は 7 で、 $6 < 7 < 21$ ですので、群 1 と群 2 の中央値に差があるといえないという判定になります。

問題

ある商品の売上増加（万円）を、条件を変えた2つのグループで測定したところ、以下の結果が得られた。2群の売上増加に差は認められるか、有意水準 5% で判定せよ。

1 群：7.2, 8.3, 5.4, 6.0, 7.3, 11.7, 10.5, 8.0, 9.1

2 群：10.1, 13.2, 7.4, 9.1, 16.2, 14.5, 6.3, 11.2, 12.4, 7.4, 12.5, 9.1, 17.0

解答

群	データ	群	データ	順位	訂正順位
1	7.2	1	5.4	1	1
1	8.3	1	6	2	2
1	5.4	2	6.3	3	3
1	6	1	7.2	4	4
1	7.3	1	7.3	5	5
1	11.7	2	7.4	6	6.5
1	10.5	2	7.4	6	6.5
1	8	1	8	8	8
1	9.1	1	8.3	9	9
2	10.1	1	9.1	10	11
2	13.2	2	9.1	10	11
2	7.4	2	9.1	10	11
2	9.1	2	10.1	13	13
2	16.2	1	10.5	14	14
2	14.5	2	11.2	15	15
2	6.3	1	11.7	16	16
2	11.2	2	12.4	17	17
2	12.4	2	12.5	18	18
2	7.4	2	13.2	19	19
2	12.5	2	14.5	20	20
2	9.1	2	16.2	21	21
2	17	2	17	22	22

- 1) 群に番号を付け、群別にデータを入力する。
- 2) データの大きさ順に並べ替える。 [データー並べ替え]
- 3) データに順位を付ける。

rank(数値,範囲,順序) 関数を利用する。

(順序：0 または省略で降順, 0 以外で昇順)

4) 同順位のものに訂正を加える。

例 6, 6 → 6.5, 6.5 [(6+7)/2] 10, 10, 10 → 11, 11, 11 [(10+11+12)/3]

5) 群別に順位合計をとる。

sumif(範囲,検索条件,合計範囲) 関数を利用する。例 sumif(C2:C23, “=1”, F2:F23)

または、再度群別に並べ直し、各群の順位合計をとる。

	データ数	順位合計
1 群	9	70
2 群	13	183

6) Wilcoxon の順位和検定数値表により、検定する。

$n_A = 9, n_B = 13$ 表より $\alpha = 0.05$ のとき → $U_1 = 73, U_2 = 134$ 両側検定

データ数の少ない 1 群の順位合計は 70 であるから、上記の範囲に入らない。

よって、有意水準 5% で差があるといえる。

問題

正規分布しない 2 群のデータで順位和を求めたところ、以下の結果を得た。それらの中央値に差があるかどうか、有意水準 5% で判定せよ。

	データ数	順位合計
1 群	30	1265
2 群	40	1220

解答

$$Z = \frac{|W - n_1(n_1 + n_2 + 1)/2| - 1/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}} = \frac{1265 - 30 \times 71/2 - 1/2}{\sqrt{30 \times 40 \times 71/12}} = 2.367629$$

$$p = 2 \cdot (1 - \text{normsdist}(2.367629)) = 0.017902 \cong 0.018$$

$p < 0.05$ より、2 群の中央値に差があるといえる。