

単回帰分析

1. 【必読】 相関係数や寄与率が 1 以上にできない理由がわかる
2. (必読) クラメールの連関係数が導出できる
3. クラメールの連関係数の値が 0、1 の条件がわかる
4. スピアマンの順位相関係数が導出できる
5. スピアマンの順位相関係数の正負の入れ替えがわかる
6. 回帰分析と実験計画法の違いがよくわかる
(繰返しデータ無しの場合)
7. 繰返しのある単回帰分析の分散分析がよくわかる
8. 回帰直線の区間推定が導出できる
9. 大波の相関、小波の相関、符号検定がよくわかる
10. 繰返しのある単回帰分析がわかる
11. 単回帰分析のテコ比がよくわかる

【必読】 相関係数や寄与率が1以上にできない理由がわかる

確かに、相関係数 $r=100$ とかないですね！

でも、何で1以上にならないかと言われても知らないし、どこにも書いていないし。。

★本記事の結論

相関係数や寄与率は、「**コーシーシュワルツの不等式に支配されているから**」

これで、「なるほど！」とわかる人は、ほばいけませんので、わかりやすく解説します。

$$(a+b)^2 = a^2 + 2ab + b^2$$

がわかれば、本記事は完璧に理解できます！

●You tube 動画もごらんください。

<https://www.youtube.com/embed/q4mq4-FJ2ik>

【1】 相関係数、寄与率

相関係数、寄与率について復習しましょう。

(1) 寄与率 R

データ群 $(x_i, y_i) (i = 1, 2, \dots, n)$ に対して、3つの平方和を定義します。

$$\bullet S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\bullet S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\bullet S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

寄与率 R は $R = \frac{S_{xy}^2}{S_{xx}S_{yy}}$ で $0 \leq R \leq 1$ が成立しますね。

(2) 相関係数 r

相関係数 r は寄与率 R の平方根ですね。

相関係数 r は $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$ で $-1 \leq r \leq 1$ が成立しますね。

相関係数、寄与率がなぜ一定の範囲に支配されているのか、を考えてみましょう。

【2】 コーシーシュワルツの不等式とその証明

(1) コーシーシュワルツの不等式

大学受験にたまに出て来る不等式です。相加相乗平均の不等式よりはマイナーですけど。

自然数 n に対して、以下の不等式が成り立つ(コーシーシュワルツの不等式)

$$(\sum_{i=1}^n a_i^2)(\sum_{i=1}^n b_i^2) \geq (\sum_{i=1}^n a_i b_i)^2$$

(2) コーシーシュワルツの不等式の証明

ここで、無理矢理感がありますが、次の2次関数を定義します。

$$2 \text{ 次関数 } f(x) = \sum_{i=1}^n (a_i x - b_i)^2$$

この2次関数は2乗和なので、**基本は $f(x) \geq 0$ です。つまり、 $y=f(x)$ と y 軸との交点の数は1か0**です。

2次関数を展開した式に書き直します。

$$f(x) = \sum_{i=1}^n (a_i x - b_i)^2 = (\sum_{i=1}^n a_i^2) x^2 - 2(\sum_{i=1}^n a_i b_i) x + 2(\sum_{i=1}^n b_i^2)$$

y 軸との交点の数をチェックする「判別式」を使います。

y 軸との交点の数が1か0なので、**判別式 $D \leq 0$ と自動的になります。**

(そうなるように2次関数を仕込みました)

$$\text{判別式 } D/4 = (\sum_{i=1}^n a_i b_i)^2 - (\sum_{i=1}^n a_i^2)(\sum_{i=1}^n b_i^2) \leq 0$$

(左辺)、(右辺) を逆にすると(コーシーシュワルツの不等式)そのものになるのがわかります。

【3】なぜ相関係数、寄与率がコーシーシュワルツの不等式に支配されるのか？

さて、相関係数、寄与率と、全く関係のない、コーシーシュワルツの不等式をつなぎます。

(1) 変数を置き換えるとわかる！

$$\bullet x_i - \bar{x} = a_i$$

$$\bullet y_i - \bar{y} = b_i$$

と置きます。

(コーシーシュワルツの不等式)は

$$(\sum_{i=1}^n a_i^2)(\sum_{i=1}^n b_i^2) \geq (\sum_{i=1}^n a_i b_i)^2 \quad \text{から}$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2 \geq (\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}))^2$$

から、

$$S_{xx} S_{yy} \geq S_{xy}^2$$

となり、すべて正ですから、

$$R = \frac{S_{xy}}{S_{xx} S_{yy}} \quad \text{で} \quad 0 \leq R \leq 1$$

が成り立ちます。

相関係数、寄与率はコーシーシュワルツの不等式に支配されている！ことがわかりますよね！

以上、相関係数や寄与率が1以上にできない理由をわかりやすく解説しました。

(必読) クラメールの連関係数が導出できる

【1】 クラメールの連関係数とは

(1) クラメールの連関係数とは

$$\bullet \text{クラメールの連関係数 } r_c = \sqrt{\frac{\chi^2}{n(k-1)}}$$

(χ^2 = 各セルについての $\frac{(\text{実測度数}-\text{期待度数})^2}{\text{期待度数}}$ 、 n : データ数、 k : 少ない方のカテゴリーの個数)

ここで、おさえるポイントは、

① 適合度の検定を使って、 χ^2 分布を使うこと

② $0 \leq r_c = \sqrt{\frac{\chi^2}{n(k-1)}} \leq 1$ なので、 r_c を相関係数みたいに使う

(2) クラメールの連関係数の計算例

よく、「クラメールの連関係数」で検索したり、教科書を読むと解き方や事例集ばかり解説していますが、

$r_c = \sqrt{\frac{\chi^2}{n(k-1)}}$ は どうやって導出したの？ 何で、こんな変な式なの？ 何で、適合度の検定 χ^2 分布なの？

と疑問に思いますよね。

(3) 適合度の検定で χ^2 分布を使う理由が説明できますか？

関連記事で確認しましょう。

関連記事 【【6】 分割表(χ^2 乗分布)に関する検定 【QC 検定®2 級対策】】

<https://qcplanets.com/qc/qc2/test-estimation6/>

大事なポイントは

1. $A = \frac{(\text{実測度数}-\text{期待度数})}{\text{期待度数}} \equiv \frac{x-\bar{x}}{\sigma}$ のように見る。
2. $\sum A^2$ を分散のように扱うので χ^2 乗分布で扱う
3. 自由度は(列-1)×(行-1) も理解しておく

ですね。ここがわからないと、クラメールの連関係数は公式暗記で終わってしまいます。。

【2】 クラメールの連関係数を導出する

(1) クラメールの連関係数を変形

$r_c = \sqrt{\frac{\chi^2}{n(k-1)}}$ 自身は重要ではなく、

$0 \leq \chi^2 \leq n(k-1)$ を証明することを考えます。

(2) 観測度数と期待度数を用意する

ここで、行×列が $J \times I$ の表を用意して、観測度数と期待度数を文字式で表現します。

★観測度数

下表のようにまとめます。

j/i	1	...	i	...	I	計
1	n_{11}	...	n_{i1}		n_{I1}	f_1
...	...	...	...	...	...	...
j	n_{1j}	...	n_{ij}		n_{Ij}	f_j
...	...	...	...	...	...	...
J	n_{1J}	...	n_{iJ}		n_{IJ}	f_J
計	g_1	...	g_i		g_I	N

★期待度数

同様に下表のようにまとめます。ここで、j 行 i 列の期待度数 E_{ij} は次式で表現できます。

$$E_{ij} = \frac{g_i}{N} \times \frac{f_j}{N} \times N = \frac{g_i f_j}{N}$$

期待度数を下表にまとめます。

j/i	1	...	i	...	I	計
1	$E_{11} = \frac{g_1 f_1}{N}$	...	$E_{i1} = \frac{g_i f_1}{N}$	...	$E_{I1} = \frac{g_I f_1}{N}$	f_1
...	...	...	...	...	...	...
j	$E_{1j} = \frac{g_1 f_j}{N}$	...	$E_{ij} = \frac{g_i f_j}{N}$	...	$E_{Ij} = \frac{g_I f_j}{N}$	f_j
...	...	...	...	...	...	...
J	$E_{1J} = \frac{g_1 f_J}{N}$	...	$E_{iJ} = \frac{g_i f_J}{N}$	...	$E_{IJ} = \frac{g_I f_J}{N}$	f_J
計	g_1	...	g_i	...	g_I	N

(3) χ^2 を計算

定義どおり χ^2 を計算します。

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(\text{実測度数} - \text{期待度数})^2}{\text{期待度数}} = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - E_{ij})^2}{E_{ij}} = (\text{式 1})$$

2 乗を展開します。

$$(\text{式 1}) = \sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{E_{ij}} - 2n_{ij} + E_{ij} \right) = (\text{式 2})$$

ここで、(式 2)の第 1 項は $\sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{E_{ij}} \right) = \sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{\frac{g_i f_j}{N}} \right) = N \sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{g_i f_j} \right)$ と変形します。

次に、(式 2)の第 2 項は $\sum_{i=1}^I \sum_{j=1}^J (-2n_{ij}) = -2N$ です。

そして、(式 2)の第 3 項は

$$\sum_{i=1}^I \sum_{j=1}^J (E_{ij}) = \sum_{i=1}^I \sum_{j=1}^J \left(\frac{g_i f_j}{N} \right) = \frac{1}{N} \sum_{i=1}^I g_i \sum_{j=1}^J f_j = \frac{1}{N} N \times N = N \text{ です。}$$

(式 2)をまとめると、

$$(\text{式 2}) = N \left(\sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{g_i f_j} \right) - 1 \right) = (\text{式 3})$$

となります。

(4) 証明のゴールを確認

で、ここで、ゴールを確認すると、

$0 \leq \chi^2 \leq n(k-1)$ を証明することを考えます。

$0 \leq \chi^2$ は明らかですね！

$$\chi^2 = \left(\sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{g_i f_j} \right) - 1 \right) \text{ から } N \Rightarrow n, \quad \sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{g_i f_j} \right) \Rightarrow k \text{ と置くと、}$$

$\chi^2 \leq n(k-1)$ の形になっていますね。

つまり、ゴールまでもう少しですね。

(5) 証明の最後まで進めると

(式 3)の $\sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{g_i f_j} \right)$ ですが、そもそも

● $\sum_{i=1}^I n_{ij} \leq f_j$ かつ ● $\sum_{j=1}^J n_{ij} \leq g_i$ ですよ。ここに注目すると
 $n_{ij} \leq f_j$ かつ $n_{ij} \leq g_i$ が共に成立しますね。

この不等式を $\sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{g_i f_j} \right)$ に代入すると

$$\sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{g_i f_j} \right) \leq \sum_{i=1}^I \sum_{j=1}^J \left(\frac{g_i \times n_{ij}}{g_i f_j} \right) = \sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}}{f_j} \right) = \sum_{j=1}^J \left(\frac{n_{1j} + n_{2j} + \dots + n_{Ij}}{f_j} \right) = \sum_{j=1}^J \left(\frac{f_j}{f_j} \right) = J$$

まとめると、

$$\sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{g_i f_j} \right) \leq J \text{ となります。}$$

同様に、

$$\sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{g_i f_j} \right) \leq \sum_{i=1}^I \sum_{j=1}^J \left(\frac{f_j \times n_{ij}}{g_i f_j} \right) = \sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}}{g_i} \right) = \sum_{i=1}^I \left(\frac{n_{i1} + n_{i2} + \dots + n_{iJ}}{g_i} \right) = \sum_{i=1}^I \left(\frac{g_i}{g_i} \right) = I$$

まとめると、

$$\sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{g_i f_j} \right) \leq I \text{ となります。}$$

つまり、共に上の 2 つの不等式が

成り立つわけですから、

$$\sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{g_i f_j} \right) \leq J \text{ かつ } \sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{g_i f_j} \right) \leq I \text{ となり、まとめると、}$$

$$\sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{g_i f_j} \right) \leq k = \min(I, J) \text{ とします。}$$

(6) 以上から

(式 3)と上の不等式の関係性を代入すると

$$(式 3) = \chi^2 = \left(\sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{g_i f_j} \right) - 1 \right) \leq n(k-1) \quad (N \rightarrow n, k = \min(I, J))$$

よって、

$0 \leq \chi^2 \leq n(k-1)$ が証明できました。

$0 \leq r_c = \sqrt{\frac{\chi^2}{n(k-1)}} \leq 1 \text{ なので、 } r_c \text{ を相関係数のように使える。}$
--

以上、「クラメールの連関係数が導出できる」を解説しました。

【1】クラメールの連関係数を導出する

本冊子、【(必読)クラメールの連関係数が導出できる】にまとめています。確認下さい。

【2】クラメールの連関係数が 1 の場合

導出過程は次の 5 つです。

- ① クラメールの連関係数が 1 となる条件式を作る
- ② 実測度数を計算
- ③ 期待度数を計算
- ④ ある文字について χ^2 を計算
- ⑤ 全体の χ^2 を計算

結果的に、 $\chi^2 = N(k - 1)$ となれば、クラメールの連関係数 $r_c = 1$ となります。

(1) 1.クラメールの連関係数が 1 となる条件式

関連記事から、次の条件式を持てきます。

$\sum_{i=1}^I \sum_{j=1}^J \left(\frac{n_{ij}^2}{g_i f_j} \right)$ から $\bullet \sum_{i=1}^I n_{ij} \leq f_j$ かつ $\bullet \sum_{j=1}^J n_{ij} \leq g_i$ が共に成り立ちます。

ここで、等号になる場合 $n_{ij} = f_j$ かつ $n_{ij} = g_i$ を考えます。この等号条件を使っていきます。

実際に式を書き出してみると

$\bullet n_{i1} + n_{i2} + \dots + n_{iJ} = g_i = n_{ij}$ $\bullet n_{1j} + n_{2j} + \dots + n_{Ij} = f_j = n_{ij}$ を同時に満たすには、

$n_{ij} = 0 (i \neq j)$ という条件になってしまいます。よって、次式のように置きます。

$$n_{ij} = \begin{cases} f_j (i = j) \text{ ただし } (I \geq J) \\ 0 (i \neq j) \end{cases}$$

(2) 2.実測度数を計算

実測度数は下表のように置けます。ここで、 $n_{ij} = f_j (I \geq J)$ としたので、 $J + 1$ から I までの値は 0 とします。

j/i	1	2	...	J	$J + 1$...	I	計
1	f_1	0	...	0	0	...	f_1
2	0	f_2		0	0	...	f_2
...	...	...	...	...	...	...	...
J	0	0	...	f_J	0	...	f_J
計	f_1	f_2	...	f_J	0	0	N

(3) 3.期待度数を計算

次に期待度数を計算します。期待度数は $\frac{f_j}{N} \times \frac{g_i}{N} \times N = \frac{f_j g_i}{N}$ に注意して解きます。下表の結果になります。

j/i	1	2	...	J	$J + 1$...	I	計
1	$\frac{f_1^2}{N}$	$\frac{f_1 f_2}{N}$	...	$\frac{f_1 f_J}{N}$	0	...	f_1
2	$\frac{f_2 f_1}{N}$	$\frac{f_2^2}{N}$		$\frac{f_2 f_J}{N}$	0	...	f_2
...	...	...	...	...	...	...	...
J	$\frac{f_J f_1}{N}$	$\frac{f_J f_2}{N}$	...	$\frac{f_J^2}{N}$	0	...	f_J
計	$g_1 = f_1$	$g_2 = f_2$	...	$g_J = f_J$	$g_{J+1} = 0$	$g_I = 0$	N

(4) 4. χ^2 を計算(その 1)

まず、 $j = 1$ についての χ^2 を計算します。実測度数と期待度数を比較しましょう。下表のとおりです。

j/i	1	2	...	J	$J + 1$	...	I	計
1(実測) f_1	f_1	0	...	0	0	...	0	f_1
1(期待) $\frac{f_1^2}{N}$	$\frac{f_1^2}{N}$	$\frac{f_1 f_2}{N}$	...	$\frac{f_1 f_J}{N}$	0	...	0	f_1

$j = 1$ についての χ^2 は

$$\chi^2 = \sum_{i=1}^I \frac{(\text{実測度数} - \text{期待度数})^2}{\text{期待度数}} = \frac{\left(f_1 - \frac{f_1^2}{N}\right)^2}{\frac{f_1^2}{N}} + \frac{\left(0 - \frac{f_1 f_2}{N}\right)^2}{\frac{f_1 f_2}{N}} + \dots + \frac{\left(0 - \frac{f_1 f_J}{N}\right)^2}{\frac{f_1 f_J}{N}}$$

$$= N \frac{f_1^2 \left(1 - \frac{f_1}{N}\right)^2}{f_1^2} + \left(\frac{f_1 f_2}{N} + \dots + \frac{f_1 f_J}{N}\right) = N \left(1 - \frac{f_1}{N}\right)^2 + \frac{f_1}{N} (f_2 + \dots + f_J) = N \left(1 - \frac{f_1}{N}\right)^2 + \frac{f_1}{N} \left(1 - \frac{f_1}{N}\right) = N - f_1 \text{ となります。}$$

($f_2 + \dots + f_J = 1 - f_1$)

(5) 5. χ^2 を計算(その 2)

同様にすべての j について χ^2 を計算すると、

$$\chi^2 = (N - f_1) + (N - f_2) + \dots + (N - f_J) = N(J - 1) \text{ となります。}$$

J は I, J の小さい方なので、 $N = n, J = k$ と置くと、

$$\chi^2 = n(k - 1) \text{ となります。}$$

よって、クラメールの連関係数 r_c は

$$r_c = \sqrt{\frac{\chi^2}{n(k-1)}} = \sqrt{\frac{n(k-1)}{n(k-1)}} = 1 \text{ と、確かに、} r_c = 1 \text{ になりましたね。面白い！}$$

【3】クラメールの連関係数が 0 の場合

(1) $\chi^2 = 0$ となる条件を探せばいい

$$\chi^2 \text{ はよく見ると } \chi^2 = (x_1 - a_1)^2 + (x_2 - a_2)^2 + \dots + (x_n - a_n)^2$$

の 2 乗和の形をしています。この 2 乗和の値を 0 にするには、

すべての i について、 $(x_i - a_i)^2 = 0$ つまり、 $x_i = a_i$ が必要。

$$\chi^2 = \sum_{i=1}^I \frac{(\text{実測度数} - \text{期待度数})^2}{\text{期待度数}} \text{ を見ると、}$$

すべての i について、**実測度数=期待度数**

となります。具体的なデータを上げると下表のような感じです。

j/i	1	2	...	I	計
1	$n_{11} = \frac{f_1^2}{N}$	$n_{21} = \frac{f_2 f_1}{N}$	...	$n_{I1} = \frac{f_I f_1}{N}$	f_1
2	$n_{12} = \frac{f_1 f_2}{N}$	$n_{22} = \frac{f_2^2}{N}$	...	$n_{I2} = \frac{f_I f_2}{N}$	f_2
...	...	...	...	...	...
J	$n_{1J} = \frac{f_1 f_J}{N}$	$n_{2J} = \frac{f_2 f_J}{N}$	...	$n_{IJ} = \frac{f_I f_J}{N}$	f_J
計	g_1	g_2	...	g_I	N

すべての i, j について、実測度数 n_{ij} と期待度数 $\frac{f_i g_j}{N}$ が等しい場合にクラメールの連関係数は 0 になります。

つまり、**実測度数がまったくばらつかないデータの場合だけ**であり、実際はそんなことは起きにくいですね。

難しい導出過程でしたが、ちゃんとできましたね！公式は導出できてから使いましょう。

以上、「クラメールの連関係数の値が 0、1 の条件がわかる」を解説しました。

スピアマンの順位相関係数が導出できる

【1】スピアマンの順位相関係数とは何か？

(1) スピアマンの順位相関係数とは何か？

2 変数をそれぞれ順位に並び替えた場合に求められる相関係数です。ピアソンの相関係数もありますが、スピアマンの順位相関係数もあります。データを下表の左から右のように順位を提示します。

No	x	y	→	x(順位)	y(順位)
1	0.15	8.05	→	1	3
2	1.2	4.05	→	2	1
3	2.08	5.77	→	3	2
4	2.42	11.2	→	4	5
5	4.82	20.17	→	5	9
6	5.93	17.21	→	6	7
7	6.15	15.22	→	7	6
8	6.5	18.38	→	8	8
9	7.32	30.59	→	9	10
10	8.45	8.99	→	10	4

(2) ピアソンの相関係数とスピアマンの順位相関係数

- ピアソンの相関係数 $r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$
- スピアマンの順位相関係数 $r = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)}$ ($d_i = x_i - y_i$)

スピアマンの順位相関係数はちょっと見かけない式ですが、導出できます。

【2】スピアマンの順位相関係数の導出の流れ

(1) 変数の準備

x, y がそれぞれ順位の変数に変わるので、それぞれ以下の変数に変わります。

- $x : 1, 2, 3, \dots, n$
- $y : 1, 2, 3, \dots, n$

(2) 解法の流れ

以下の通りに計算していきます。

- ① 最初に平均 $\bar{x}, \bar{y}$ を計算
- ② 平方和 S_{xx}, S_{yy}, S_{xy} を計算

【3】スピアマンの順位相関係数を導出(2通り表現できる)

(1) 最初に平均 $\bar{x}, \bar{y}$ を計算

- $\bar{x} = E[x] = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2}$
- $\bar{y} = E[y] = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2}$

と同じとなります。

(2) 平方和 S_{xx}, S_{yy}, S_{xy} を計算

- $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + \bar{x}^2 \sum_{i=1}^n 1$
 $= \frac{1}{6} n(n+1)(2n+1) - 2 \times \frac{n+1}{2} \times \frac{n(n+1)}{2} + \frac{(n+1)^2}{4} \times n = \frac{n(n^2-1)}{12}$
- $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{n(n^2-1)}{12} = S_{xx}$

● S_{xy} は工夫します。

$\sum_{i=1}^n (x_i - y_i)^2 = \sum_{i=1}^n d_i^2$ を考えます。

$$\begin{aligned}\sum_{i=1}^n (x_i - y_i)^2 &= \sum_{i=1}^n ((x_i - \bar{x}) - (y_i - \bar{y}))^2 = \sum_{i=1}^n (x_i - \bar{x})^2 - 2 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= S_{xx} - 2S_{xy} + S_{yy}\end{aligned}$$

まとめると、

$$\sum_{i=1}^n d_i^2 = S_{xx} - 2S_{xy} + S_{yy} \quad \text{から}$$

$$S_{xy} = \frac{1}{2}(S_{xx} + S_{yy} - \sum_{i=1}^n d_i^2)$$

となります。

$$S_{yy} = S_{xx} = S \text{ とおくと、}$$

スピアマンの順位相関係数

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\frac{1}{2}(2S - \sum_{i=1}^n d_i^2)}{S} = 1 - \frac{1}{2} \frac{\sum_{i=1}^n d_i^2}{S} = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

とシンプルな公式ができます。

難しい導出過程でしたが、ちゃんとできましたね！公式は導出できてから使いましょう。

以上、「スピアマンの順位相関係数が導出できる」を解説しました。

スピアマンの順位相関係数の正負の入れ替えがわかる

【1】スピアマンの順位相関係数の正負が変わる条件

(1) 正負の入れ替え方

- x または y のどちらかの順位を入れ替えるとスピアマンの順位相関係数の正負は入れ替わる。
- x, y の両方の順位を入れ替えるとスピアマンの順位相関係数の正負は2回入れ替わるので、もとの正負に戻る。

(2) 正負の入れ替え事例

実際にやってみましょう。下表のように4つ条件を作ります。

- 条件1：元データ
 条件2： x だけ順位を入れ替えた場合
 条件3： y だけ順位を入れ替えた場合
 条件4： x, y 両方順位を入れ替えた場合

	条件1		条件2		条件3		条件4	
No	x	y	x	y	x	y	x	y
1	1	3	10	3	1	8	10	8
2	2	1	9	1	2	10	9	10
3	3	2	8	2	3	9	8	9
4	4	5	7	5	4	6	7	6
5	5	9	6	9	5	2	6	2
6	6	7	5	7	6	4	5	4
7	7	6	4	6	7	5	4	5
8	8	8	3	8	8	3	3	3
9	9	10	2	10	9	1	2	1
10	10	4	1	4	10	7	1	7

黄色マーカー部分が順位が入れ替わったところです。

(3) 各条件のスピアマンの順位相関係数 r を計算

実際に、各条件のスピアマンの順位相関係数 r を計算すると下表のようになります。

	条件1	条件2	条件3	条件4
S_{xx}	82.5	82.5	82.5	82.5
S_{yy}	82.5	82.5	82.5	82.5
S_{yx}	51.5	-51.5	-51.5	51.5
r	0.624	-0.624	-0.624	0.624

- x または y のどちらかの順位を入れ替えるとスピアマンの順位相関係数の正負は入れ替わる。
- x, y の両方の順位を入れ替えるとスピアマンの順位相関係数の正負は2回入れ替わるので、もとの正負に戻る。

確かに、正負が入れ替わっていますよね。でも、なぜそうなるか？わかりますか？

【2】スピアマンの順位相関係数の正負が入れ替わる理由

大事なのは、ピアソンの相関係数の式からスピアマンの順位相関係数の性質が導出できます！スピアマンの順位相関係数のための公式暗記は一切不要！導出過程を理解しましょう！

(1) 元の条件における平方和の式を立てる

正負が入れ替わる理由を数式で説明します。もともとの平方和は、

- $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$
- $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$
- $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$

ですよね。

入れ替えとは、
 x が $(n+1)-x$ 、
 y が $(n+1)-y$ 、
 に変換することです。(例：1,2,3,4,5,6,7,8,9,10 → 10,9,8,7,6,5,4,3,2,1)

文字を入れ替えた場合の平方和に「'」をつけて計算します。

(2) (条件 2) x だけが入れ替わる場合

平方和は次のように式が変わります。

- $S_{xx}' = \sum_{i=1}^n (((n+1)-x_i) - ((n+1)-\bar{x}))^2 = \sum_{i=1}^n (-x_i + \bar{x})^2 = S_{xx}$
 - $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$
 - $S_{xy}' = \sum_{i=1}^n (((n+1)-x_i) - ((n+1)-\bar{x}))(y_i - \bar{y}) = \sum_{i=1}^n (-x_i + \bar{x})(y_i - \bar{y}) = -S_{xy}$
- と元の S_{xy} と正負が入れ替わります。

これが、スピアマンの順位相関係数 r の正負が入れ替わる理由ですね。

$$r' = \frac{S_{xy}'}{\sqrt{S_{xx}'S_{yy}'}} = -\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = -r$$

なるほど、よくわかりますね！

(3) (条件 3) y だけが入れ替わる場合

平方和は次のように式が変わります。

- $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$
- $S_{yy}' = \sum_{i=1}^n (((n+1)-y_i) - ((n+1)-\bar{y}))^2 = \sum_{i=1}^n (-y_i + \bar{y})^2 = S_{yy}$
- $S_{xy}' = \sum_{i=1}^n (x_i - \bar{x})(((n+1)-y_i) - ((n+1)-\bar{y})) = \sum_{i=1}^n (x_i - \bar{x})(-y_i + \bar{y}) = -S_{xy}$

これが、スピアマンの順位相関係数 r の正負が入れ替わる理由ですね。

なるほど、よくわかりますね！

(4) (条件 4) x, y 両方が入れ替わる場合

平方和は次のように式が変わります。

- $S_{xx}' = S_{xx}$
- $S_{yy}' = S_{yy}$
- $S_{xy}' = \sum_{i=1}^n (((n+1)-x_i) - ((n+1)-\bar{x}))(((n+1)-y_i) - ((n+1)-\bar{y})) = \sum_{i=1}^n (-x_i + \bar{x})(-y_i + \bar{y}) = S_{xy}$

これが、スピアマンの順位相関係数 r の正負が元に戻る理由ですね。

(5) データを再度見て確認しよう！

データを再掲すると、計算通りの結果になっていますよね。

-	条件1	条件2	条件3	条件4
S_{xx}	82.5	82.5	82.5	82.5
S_{yy}	82.5	82.5	82.5	82.5
S_{yx}	51.5	-51.5	-51.5	51.5
r	0.624	-0.624	-0.624	0.624

ちゃんと、説明が付きましてね！

大事なものは、ピアソンの相関係数の式からスピアマンの順位相関係数の性質が導出できます！スピアマンの順位相関係数のための公式暗記は一切不要！導出過程を理解しましょう！

以上、「スピアマンの順位相関係数の正負の入れ替えがわかる」を解説しました。

確かに、同じ分散分析しますが、回帰分析と実験計画法では何が違うのか？何をそれぞれ見ているのか？が気になりますね。（気になってほしいです！）

なので、解説します！

- (i)モデル式である「データの構造式」を立てて
- (ii)データの構造式から「平方和の分解」を確認して
- (iii)分散分析して
- (iv)分散分析の結果を比較します！

【1】単回帰分析による分散分析

(1) データの用意

例えば、下表のようなデータを用意します。各値は次の通りです（計算してみてください）。

No	x	y
1	0.15	8.05
2	1.2	4.05
3	2.08	5.77
4	2.42	11.2
5	4.82	20.17
6	5.93	17.21
7	6.15	15.22
8	6.5	18.38
9	7.32	30.59
10	8.45	8.99
合計	45.02	139.63

- 平方和 $S_{xx} = 72.42$
- 平方和 $S_{yy} = 579.34$
- 平方和 $S_{xy} = 128.79$

より、

● 相関係数 $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.629$

● 回帰平方和 $S_R = \frac{S_{xy}^2}{S_{xx}} = 229.04$

● 残差平方和 $S_{er} = S_T - S_R = 350.30$

● 総平方和 $S_T = S_{yy} = 579.34$

(2) 単回帰分析による分散分析

各平方和が計算出来たので、分散分析は下表のとおりになります。

回帰	S	Φ	V
R	229.04	1	229.04
er	350.3	8	43.79
T	579.34	9	-

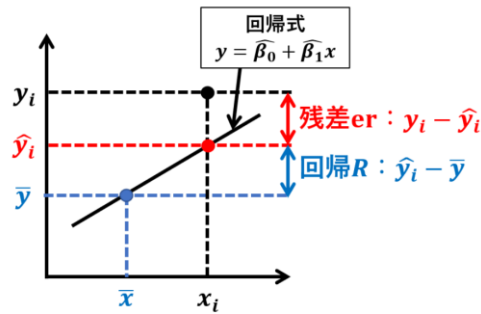
【2】単回帰分析による平方和の分解

(1) データの構造式

単回帰分析のデータの構造式を書いてみましょう。文字式を以下のように定義します。

- データ $\rightarrow (x_i, y_i)$
- 平均 $\rightarrow (\bar{x}, \bar{y})$
- 回帰直線上のデータ $\rightarrow (x_i, \hat{y}_i)$

下図のとおりです。



ポイントは、データ x_i と回帰直線上のデータ $\rightarrow x_i$ は同じである点です。平方和の分解で必要になってきます。

単回帰分析のデータの構造式は、

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

となりますね。上図と見ながら確認しましょう。なお、データの構造式を見ると

- 全体： $(y_i - \bar{y})$
- 回帰： $(\hat{y}_i - \bar{y})$
- 残差： $(y_i - \hat{y}_i)$

の成分に分けることができますね。これが分散分析できる理由になります。

(2) 平方和の分解

実際に、分散分析するときは、

$$\text{総平方和 } S_T = \text{回帰平方和 } S_R + \text{残差平方和 } S_{er} \quad (*)$$

と分けますが、式で書くと

- 総平方和： $S_T = \sum_{i=1}^n (y_i - \bar{y})^2$
- 回帰平方和： $S_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$
- 残差平方和： $S_{er} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

となりますね。これが回帰分析で分散分析できる理由です。(*)を証明してみましょう。

次に同じ分散分析でも実験計画法で考えてみましょう。

【3】 繰返しのない一元配置実験による分散分析

(1) データの用意

①の回帰分析と同じデータを用意しますが、実験計画法用に表を作り直します。

No	x	y	因子	y
1	0.15	8.05	A ₁	8.05
2	1.2	4.05	A ₂	4.05
3	2.08	5.77	A ₃	5.77
4	2.42	11.2	A ₄	11.2
5	4.82	20.17	A ₅	20.17
6	5.93	17.21	A ₆	17.21
7	6.15	15.22	A ₇	15.22
8	6.5	18.38	A ₈	18.38
9	7.32	30.59	A ₉	30.59
10	8.45	8.99	A ₁₀	8.99
合計	45.02	139.63	合計	139.63

実は、繰返し実験のない一元配置実験の表になります。

(2) データの構造式から分散分析へ

繰返し実験のない一元配置実験のデータの構造式は

$$(y_i - \bar{y}) = (y_i - \bar{y})$$

となり、主効果が一切なく、総平方和＝残差平方和という変なパターンになります。分散分析表を書くと下表になります。

	平方和S	自由度Φ	平均平方V
主効果	－	－	－
残差e	579.34	9	64.37
合計T	579.34	9	－

同じデータで回帰分析と実験計画法を使って分散分析しました。ここから両者を比較しましょう。

【4】平方和を分解して回帰分析と実験計画法を比較

(1) 分散分析結果を比較

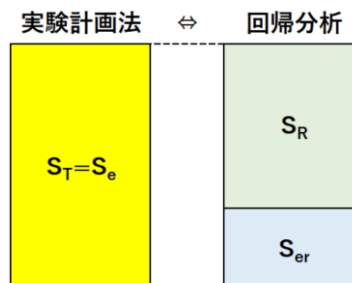
実験計画法	平方和S	自由度Φ	平均平方V	回帰	平方和S	自由度Φ	平均平方V
主効果	－	－	－	回帰	229.04	1	229.04
残差e	579.34	9	64.37	残差er	350.3	8	43.79
合計T	579.34	9	－	合計T	579.34	9	－

平方和に注目すると

●総平方和＝回帰平方和＋回帰残差平方和（回帰分析）

●総平方和＝主効果平方和＋残差平方和（実験計画法）

に分割できる点です。



もう少しモデルが複雑にすると、主効果の一部が回帰の平方和に分割できることがわかります。これも関連記事に上げていきます。

(2) 回帰分析と実験計画法の違い

データの構造式で比較すると

●実験計画法： $(y_i - \bar{y}) = (y_i - \bar{y})$

●回帰分析： $(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$

として、回帰成分で総平方和を分割しているイメージがわかりますね。

回帰分析と実験計画法は、分散分析するので、総平方和からどの成分がどれくらい分解されるかを意識して計算しましょう。

以上、「回帰分析と実験計画法の違いがよくわかる（繰返しデータ無しの場合）がよくわかる」を解説しました。

確かに、同じ分散分析しますが、回帰分析と実験計画法では何が違うのか？何をそれぞれ見ているのか？が気になりますね。なので、解説します！

- (i)モデル式である「データの構造式」を立てて
- (ii)データの構造式から「平方和の分解」を確認して
- (iii)分散分析して
- (iv)分散分析の結果を比較します！

繰返しの無い場合については本冊子【回帰分析と実験計画法の違いがよくわかる（繰返しデータ無しの場合）】で解説しています。

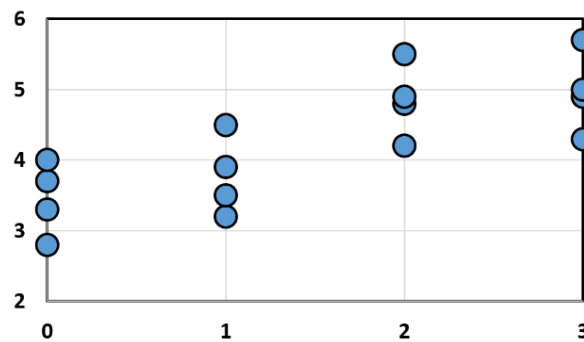
【1】単回帰分析による分散分析

(1) データの用意

例えば、下表のようなデータを用意します。

i	j	x_i	y_{ij}	i	j	x_i	y_{ij}	i	j	x_i	y_{ij}	i	j	x_i	y_{ij}
1	1	0	3.3	2	1	1	4.5	3	1	2	4.8	4	1	3	4.9
1	2	0	2.8	2	2	1	3.9	3	2	2	4.2	4	2	3	5.7
1	3	0	3.7	2	3	1	3.2	3	3	2	5.5	4	3	3	5.0
1	4	0	4.0	2	4	1	3.5	3	4	2	4.9	4	4	3	4.3
-	合計	24	68.2	-	-	-	-	-	-	-	-	-	-	-	-
-	平均	1.5	4.63	-	-	-	-	-	-	-	-	-	-	-	-

グラフを描くとこんな感じです。



確かに、繰返しがあるデータですね。i,j で区分していますが、気にせず全 16 個のデータについて寄与率、相関係数を計算します！

なお、各値は次の通りです（計算してみてください）。

- 平方和 $S_{xx}=20$
- 平方和 $S_{yy}=10.60$
- 平方和 $S_{xy}=11.3$
- 寄与率 $R=\frac{S_{xy}^2}{S_{xx}S_{yy}}=0.776$
- 回帰平方和 $S_R=\frac{S_{xy}^2}{S_{xx}}=6.38$
- 残差平方和 $S_{er}=S_T - S_R=4.21$
- 総平方和 $S_T=S_{yy}=10.60$

(2) 単回帰分析による分散分析

各平方和が計算出来たので、分散分析は下表のとおりになります。

-	S	ϕ
R	6.38	1
er	4.22	14
T	10.60	15

次に、同じデータを実験計画法の分散分析をやってみましょう。

【2】一元配置実験による分散分析

(1) よくみると一元配置実験である

繰返しのある単回帰分析のデータは、一元配置実験と同じなんです。実験計画法っぽく表を作り変えます。

A	データ			
A0	3.3	2.8	3.7	4
A1	4.5	3.9	3.2	3.5
A2	4.8	4.2	5.5	4.9
A3	4.9	5.7	5	4.3

こう書くと、単純な一元配置実験ですね。

同じデータでも

- 一元配置実験はQC検定@2級レベル
- 繰返しのある単回帰分析はQC検定@級レベル

別々とせず、両者の違いを理解しながら、まとめて勉強しましょう！その方が速い！

(2) データの構造式

データ構造式はQCプラネッツでいっぱい書いていますので関連記事で復習しましょう。

【関連記事】【簡単】データの構造式で実験計画法がわかる(必読)

<https://qcplanets.com/method/doe/data-structure/>

一元配置実験のデータ構造式は、下のとおりです。

$$(y_{ij} - \bar{y}) = (\bar{y}_i - \bar{y})(\text{主効果}) + (y_{ij} - \bar{y}_i)(\text{残差})$$

(3) 一元配置実験による分散分析

実験計画法による分散分析すると次の結果になります。

単回帰分析	S	Φ	実験計画法	S	Φ
R	6.38	1	A	7.00	3
er	4.22	14	e	3.60	12
T	10.60	15	T	10.60	15

【3】分散分析の比較(回帰分析 vs 実験計画法)

元データは同じですが、次の2つができました。比較してみましょう。

- 繰返しのある単回帰分析による分散分析
- 一元配置実験による分散分析

単回帰分析	S	Φ	実験計画法	S	Φ
R	6.38	1	A	7.00	3
er	4.22	14	e	3.60	12
T	10.60	15	T	10.60	15

となり、合計の平方和と自由度は同じですが、回帰R、主効果A、残差eとerでそれぞれ値が違いますね。単回帰分析では、両者の違いを埋めるために当てはまりの悪さ lof を使います。

$$S_T = S_R + S_{er} = S_A + S_e \text{ より}$$

$$S_{lof} = S_A - S_R = S_{er} - S_e \text{ から求めます。}$$

自由度も同様に、 $S \rightarrow \phi$ に変えて計算しましょう。再度、分散分析表で両手法を比較しましょう。

単回帰分析		S	Φ	実験計画法	S	Φ
R		6.38	1	A	7.00	3
er	lof	0.62	2			
er	e	3.60	12	e	3.60	12
T		10.60	15	T	10.60	15

繰返しありの単回帰分析で分散分析表と当てはまりの悪さを早くマスターしたいなら、
一元配置実験を持ってくると良いです！

繰返しありの単回帰分析で難しい分散分析が簡単に理解できましたね！

【4】（必読！）データの構造式と平方和の分解

単回帰分析と実験計画法から分散分析しました。その元になる理論も解説します。平方和の分解をひたすら計算します。

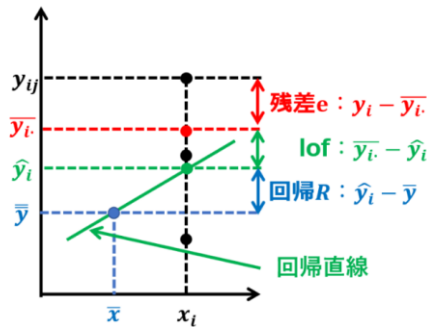
(1) データの構造式

単回帰分析のデータの構造式

$(y_{ij} - \bar{y}) = (\hat{y}_i - \bar{y})(\text{回帰 } R) + (y_{ij} - \hat{y}_i)(\text{残差})$ 単回帰分析のデータの構造式

$(y_{ij} - \bar{y}) = (\bar{y}_i - \bar{y})(\text{主効果}) + (y_{ij} - \bar{y}_i)(\text{残差})$ 実験計画法・一元配置実験のデータの構造式

図で関係性を確認します。



(2) 平方和の分解

上の図から、単回帰分析と実験計画法の両者のデータの構造式をまとめると次の式ができます。

$(y_{ij} - \bar{y}) = (\hat{y}_i - \bar{y})(\text{回帰 } R) + (\bar{y}_i - \hat{y}_i)(\text{lof}) + (y_{ij} - \bar{y}_i)(\text{残差 } e)$

ここからそれぞれの平方和を分解することができます。

平方和を求める式を作ると

$$\sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y})^2 = \sum_{i=1}^a \sum_{j=1}^b (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^a \sum_{j=1}^b (\bar{y}_i - \hat{y}_i)^2 + \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y}_i)^2$$

$$S_T = S_R + S_{\text{lof}} + S_{\text{er}}$$

の関係式になります。（確かめてみてください。）

なお、計算結果はここにありますのでご確認ください。

リンク：<https://qcplanets.com/wp-content/uploads/2023/01/shoumei.pdf>

ここまで、読んでいただくと、回帰分析と実験計画法の両方の分散分析が自由自在に使いこなせるはずです。

以上、「繰返しのある単回帰分析の分散分析がよくわかる」を解説しました。

【1】回帰直線の推定区間とは

(1) 回帰直線の求め方は、関連記事に書いています。

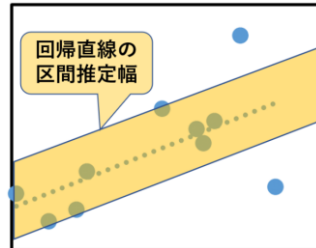
【関連記事】回帰分析と相関係数をマスターする

<https://qcplanets.com/method/regression/basic/>

回帰直線が導出できることを前提として、区間推定を解説します。

(2) 回帰直線の区間推定とは

下図のように、回帰直線には、区間推定の領域があります。



(3) 回帰直線の係数の推定区間を導出

回帰直線に区間推定の領域がある理由は

実データ y_i はランダムな誤差とする残差 ε_i が入り、ばらつきを与えるので、
回帰直線の傾き a と y 切片 b もばらつくため、区間推定を考える必要があります。

(4) 回帰直線の係数の推定区間の式が難しい

回帰直線の傾き a と y 切片 b の区間推定の式が出来れば OK ですが、

● 傾き a は、平均 a , 分散 $\frac{\sigma^2}{S_{xx}}$ の正規分布に従い、

● y 切片 b は、平均 b , 分散が $\sigma^2 \left(\frac{1}{n} + \frac{x^2}{S_{xx}} \right)$ の正規分布に従うのですが、

どうやってこの式になるのか？疑問ですね！

導出できない公式は使うな！式の導出をちゃんと理解しよう！

ということで、本記事で導出していきます。なので、以下の流れで解いていきます。

- ① まず、導出の途中過程で必要な式を事前に計算
- ② 回帰直線の傾き a の期待値と分散を導出
- ③ 回帰直線の y 切片 b の期待値と分散を導出
- ④ 回帰直線の区間推定領域を図示

【2】【重要】回帰直線の区間推定導出のポイント

回帰直線の傾き a と y 切片 b の期待値と分散を導出する際に、最も意識してほしいのは、

誤差を含むのは y_i だけ。 x_i だけの式は定数扱いで変形していくこと。

分散 $V[a]$ の式の $[a]$ の中から外にさらっと出す式変形が多々できますが、

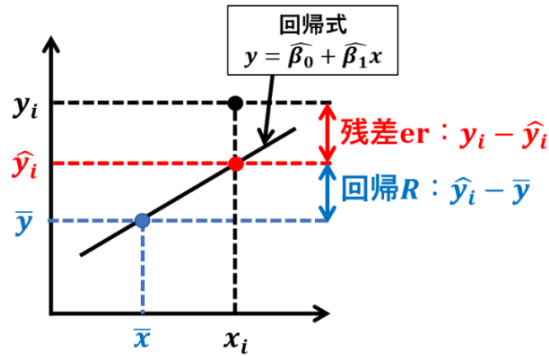
変数は y に関わるものとしてみてください。ここが抜けると式変形が完全に理解できません。

【3】導出過程に必要な値を事前に計算

(1) 各値の定義

平均 $\bar{x}, \bar{y}$ 、回帰直線に乗る点 $\hat{x}_i, \hat{y}_i$ の位置関係を下図にあります。分散分析も含めた関連記事で復習しましょう。

(本冊子の【繰返しのある単回帰分析の分散分析がよくわかる】)



事前に導出した値は以下の通りです。

1. $V[y_i] = \sigma^2$
2. $V[\bar{y}] = \frac{\sigma^2}{n}$
3. 共分散の性質で、 X_i と Y_j が互いに独立ならば、

$$\text{Cov}(X_1 + X_2, Y_1 + Y_2) = \text{Cov}(X_1, Y_1) + \text{Cov}(X_1, Y_2) + \text{Cov}(X_2, Y_1) + \text{Cov}(X_2, Y_2)$$
4. σ^2 は計測できないので、 $\sigma^2 = \text{Ve} = \frac{S_e}{\phi_e (= n-2)}$ を使う

さっと導出します。

(1) $V[y_i] = \sigma^2$ の導出

データ y_i は、 x_i 、回帰直線の傾き a 、 y 切片 b と正規分布 $N(0, \sigma^2)$ に従う(と仮定する)誤差 ε_i を使って、

$$y_i = ax_i + b + \varepsilon_i \quad \text{と表現できます。}$$

誤差 ε_i のばらつきによって、回帰直線の区間推定幅があると考えるので、次式とします。

$$V[y_i] = V[ax_i + b + \varepsilon_i] = V[\varepsilon_i] = \sigma^2$$

(2) $V[\bar{y}] = \frac{\sigma^2}{n}$ の導出

$\bar{y} = \frac{1}{n}(y_1 + y_2 + \dots + y_n)$ より、

$$V[\bar{y}] = V\left[\frac{1}{n}(y_1 + y_2 + \dots + y_n)\right] = \left[V\left(\frac{y_1}{n}\right) + V\left(\frac{y_2}{n}\right) + \dots + V\left(\frac{y_n}{n}\right)\right] = n V\left(\frac{y_i}{n}\right) = n \frac{1}{n^2} \sigma^2 = \frac{\sigma^2}{n}$$

(3) 共分散の性質で、 X_i と Y_j が互いに独立ならば、

$$\text{Cov}(X_1 + X_2, Y_1 + Y_2) = \text{Cov}(X_1, Y_1) + \text{Cov}(X_1, Y_2) + \text{Cov}(X_2, Y_1) + \text{Cov}(X_2, Y_2)$$

(4) σ^2 は計測できないので、 $\sigma^2 = \text{Ve} = \frac{S_e}{\phi_e (= n-2)}$ を使う

【4】 回帰直線の傾き a の期待値と分散を導出

(1) 求めたい値を再度確認

- $E[a] = a$
 - $V[a] = \frac{\sigma^2}{S_{xx}}$
 - $E[b] = b$
 - $V[b] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$
- を計算します。

(1) 回帰直線の傾き a の期待値 $E[a]$ を導出

回帰直線の傾き $a = \frac{S_{xy}}{S_{xx}}$ です。分母分子の関係式を考えます。

① S_{xy} を変形

特に y が変数なので

$$\begin{aligned} S_{xy} &= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})y_i - \sum_{i=1}^n (x_i - \bar{x})\bar{y} = \sum_{i=1}^n (x_i - \bar{x})y_i - \bar{y} \sum_{i=1}^n (x_i - \bar{x}) \\ &= \sum_{i=1}^n (x_i - \bar{x})y_i - \bar{y} \times 0 = \sum_{i=1}^n (x_i - \bar{x})y_i \quad (\text{マーカの式は後でも使います。}) \end{aligned}$$

② 期待値 $E[X]$ を使って変形

$y_i = ax_i + b + \varepsilon_i$ より

$$\begin{aligned} E[S_{xy}] &= E[\sum_{i=1}^n (x_i - \bar{x})y_i] = E[\sum_{i=1}^n (x_i - \bar{x})(ax_i + b + \varepsilon_i)] \\ &= a E[\sum_{i=1}^n (x_i - \bar{x})x_i] + b E[\sum_{i=1}^n (x_i - \bar{x})] + E[\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i] \end{aligned}$$

ここで、第2項において、 $E[\sum_{i=1}^n (x_i - \bar{x})] = 0$ なので、第2項は0です。

次に第3項において、 ε_i は y_i 方向についてのばらつきなので、 ε_i と $x_i - \bar{x}$ とは独立です。
独立な場合は期待値には便利な性質があり、 $E[XY] = E[X]E[Y]$ があります。これを使うと、
 $E[\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i] = E[\sum_{i=1}^n (x_i - \bar{x})] \times E[\sum_{i=1}^n \varepsilon_i] = 0 \times E[\sum_{i=1}^n \varepsilon_i] = 0$ となります。
($E[\sum_{i=1}^n (x_i - \bar{x})] = 0$ より)

まとめると、

$E[S_{xy}] = a E[\sum_{i=1}^n (x_i - \bar{x})x_i]$
--

さらに、 $\sum_{i=1}^n (x_i - \bar{x})$ が0なら式を加えても値は変わらないので、あえて

$$\begin{aligned} E[S_{xy}] &= a E[\sum_{i=1}^n (x_i - \bar{x})x_i] = a E[\sum_{i=1}^n (x_i - \bar{x})x_i] - a E[\sum_{i=1}^n (x_i - \bar{x})\bar{x}] \\ &= a E[\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})] = a E[\sum_{i=1}^n (x_i - \bar{x})^2] = a E[S_{xx}] \end{aligned}$$

となります。

x は定数扱いなので、 $E[S_{xx}]$ の中身はそのまま取り出せて、 $E[S_{xx}] = S_{xx}$ です。

よって、期待値 $E[a]$ は $E[a] = E[\frac{S_{xy}}{S_{xx}}] = a$ となります。 当たり前だけど、途中経過が難しいし、期待値の性質を使った式変形の難しさを思い知らされますね！
--

(2) 回帰直線の傾き a の分散 $V[a]$ を導出

期待値の計算で難しさをわかったところで、分散も解きましょう。

$V[a] = V[\frac{S_{xy}}{S_{xx}}]$ ですから、変数 y を含む $V[S_{xy}]$ の計算が必要です。

① 分散 $V[S_{xy}]$ の計算

先ほどの計算で、以下の3点を意識して分散 $V[S_{xy}]$ を計算します。

- | |
|---|
| <ol style="list-style-type: none">1. $S_{xy} = \sum_{k=1}^n (x_i - \bar{x})y_i$2. $y_i = ax_i + b + \varepsilon_i$3. 変数 x_i, y_i は定数扱い |
|---|

$$V[S_{xy}] = V[\sum_{k=1}^n (x_i - \bar{x}) y_i] = V[\sum_{k=1}^n (x_i - \bar{x})(ax_i + b)] + V[\sum_{k=1}^n (x_i - \bar{x}) \varepsilon_i] = (**)$$

ここで、 x についての値は定数と考えるので、

$$\bullet V[\sum_{k=1}^n (x_i - \bar{x})(ax_i + b)] = 0$$

$$\bullet V[\sum_{k=1}^n (x_i - \bar{x}) \varepsilon_i] = \sum_{k=1}^n (x_i - \bar{x})^2 V[\varepsilon_i]$$

となります。

$$V[S_{xy}] = \sum_{k=1}^n (x_i - \bar{x})^2 V[\varepsilon_i] = \sum_{k=1}^n (x_i - \bar{x})^2 \sigma^2 = \sigma^2 S_{xx}$$

よって、分散 $V[a]$ は

$$V[a] = V[\frac{S_{xy}}{S_{xx}}] = \frac{1}{S_{xx}^2} V[S_{xy}] = \frac{1}{S_{xx}^2} \sigma^2 S_{xx} = \frac{\sigma^2}{S_{xx}}$$

(S_{xx} は定数扱いなので、 V の外には2乗して出します。)

結果はシンプルですが、計算は結構難しいですね。

傾き a を使って、期待値、分散の計算が慣れてきましたので
 y 切片 b の期待値、分散の計算をしましょう。

【5】 回帰直線の y 切片 b の期待値と分散を導出

(1) 求めたい値を再度確認

a, b の期待値と分散が導出できれば本記事はOKです。

$$\bullet E[a] = a$$

$$\bullet V[a] = \frac{\sigma^2}{S_{xx}}$$

$$\bullet E[b] = b$$

$$\bullet V[b] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$

を計算します。 $E[b]$ と $V[b]$ を解きましょう。

(2) 回帰直線の y 切片 b の期待値と分散を導出

$$\text{回帰直線の } y \text{ 切片 } b = -\frac{S_{xy}}{S_{xx}} \bar{x} + \bar{y} = -a\bar{x} + \bar{y} \text{ ですね。}$$

期待値 $E[b]$ は

$$E[b] = E[-a\bar{x} + \bar{y}] = -E[a] E[\bar{x}] + E[\bar{y}]$$

となり、それぞれの期待値は以下の値を使うと

$$\bullet E[a] = a \text{ (【4】で導出)}$$

$$\bullet E[\bar{x}] = \bar{x} \text{ (定数なのでそのまま期待値 } E[\] \text{ の外に出る)}$$

$$\bullet E[\bar{y}] = \bar{y} \text{ (定数なのでそのまま期待値 } E[\] \text{ の外に出る)}$$

$$E[b] = -a\bar{x} + \bar{y} = b$$

となります。しっかり計算した結果、当たり前の結果になりましたね。

(3) 回帰直線の y 切片 b の分散 $V[b]$ を導出

$$V[b] = V[-a\bar{x} + \bar{y}]$$

ここで、

$$\bullet X, Y \text{ が互いに独立なら } V[X+Y] = V[X] + V[Y] \text{ で}$$

$$\bullet \text{独立でないなら、} V[X+Y] = V[X] + V[Y] + 2\text{Cov}(X, Y)$$

という、共分散 $\text{Cov}(X, Y)$ が出て来ますよね！

結論から言えば $-a\bar{x} + \bar{y}$
 $-a\bar{x}$ と $\bar{y}$ は互いに独立なので、
 $V[b]=V[-a\bar{x} + \bar{y}]=V[-a\bar{x}]+V[\bar{y}]$
 と分離でき、 $\text{Cov}(-a\bar{x}, \bar{y})=0$ です。

$V[b]=V[-a\bar{x} + \bar{y}]=V[-a\bar{x}]+V[\bar{y}]=(-1)^2\bar{x}^2 V[a]+V[\bar{y}]$
 とします。

● $V[a]=\frac{\sigma^2}{S_{xx}}$ (【4】で導出済。)

● $V[\bar{y}]=\frac{\sigma^2}{n}$ (【3】で導出済。)

を使います。

よって、

$V[b]=\bar{x}^2 V[a]+V[\bar{y}]=\bar{x}^2 \frac{\sigma^2}{S_{xx}} + \frac{\sigma^2}{n} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$ となります。

(4) 共分散 $\text{Cov}(-a\bar{x}, \bar{y})=0$ を確認

$\text{Cov}(-a\bar{x}, \bar{y})= \text{Cov}\left(-\frac{S_{xy}}{S_{xx}}\bar{x}, \bar{y}\right)=\frac{\bar{x}}{S_{xx}}\text{Cov}(S_{xy}, \bar{y})$

($\bar{x}$ と S_{xx} は定数扱いなので、Covの外に出します。)

次に、 S_{xy} と $\bar{y}$ を展開して、Covの中の式を変形します。

● $S_{xy}= E[\sum_{i=1}^n (x_i - \bar{x})y_i]$ (【3】で解説済です)

● $\bar{y}=\frac{1}{n}\sum_{i=1}^n y_i$ (【2】で解説済です)

と【3】のところで事前に紹介しましたが、
 共分散の性質で、 X_i と Y_j が互いに独立ならば、

● $\text{Cov}(X_1 + X_2, Y_1 + Y_2)= \text{Cov}(X_1, Y_1)+ \text{Cov}(X_1, Y_2)+ \text{Cov}(X_2, Y_1)+ \text{Cov}(X_2, Y_2)$
 となります。

以上の性質を活用すると、

$\text{Cov}(-a\bar{x}, \bar{y})=\frac{\bar{x}}{S_{xx}}\text{Cov}(S_{xy}, \bar{y})=\text{Cov}((x_1 - \bar{x})y_1, \frac{y_1}{n})+ \text{Cov}((x_2 - \bar{x})y_2, \frac{y_2}{n})+\dots+ \text{Cov}((x_n - \bar{x})y_n, \frac{y_n}{n})$

となり、 x, n は定数扱いとしてCovの外に出します。

$=\frac{x_1 - \bar{x}}{n} \text{Cov}(y_1, y_1)+\frac{x_2 - \bar{x}}{n} \text{Cov}(y_2, y_2)+\dots+\frac{x_n - \bar{x}}{n} \text{Cov}(y_n, y_n)=\sum_{i=1}^n \frac{x_i - \bar{x}}{n} \sigma^2$ (Cov(X,X)=V[X])

となりますが、

$\sum_{i=1}^n \frac{x_i - \bar{x}}{n}=0$ なので、共分散Covは0になります。

よって共分散 $\text{Cov}(-a\bar{x}, \bar{y})=0$

【6】 回帰直線の区間推定領域を図示

(1) 求めたい値を再度確認

まとめると、

● $E[a] = a$

● $V[a] = \frac{\sigma^2}{S_{xx}}$

● $E[b] = b$

● $V[b] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$

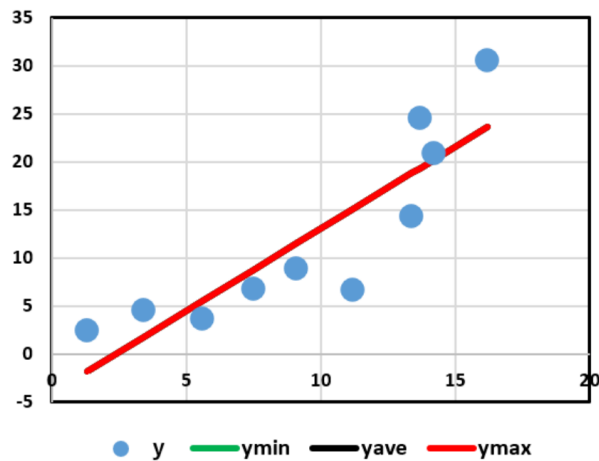
では、実際に値を使って、回帰直線の区間推定領域を計算してグラフ化しましょう。

(2) データを用意

例えば、次の値を用意します。ついでに平方和もグラフも出します。

No	x	y	A= (x - $\bar{x}$)	B= (y - $\bar{y}$)	A ²	B ²	AB
1	1.3	2.4	-8.26	-9.88	68.23	97.61	81.61
2	3.4	4.5	-6.16	-7.78	37.95	60.53	47.92
3	5.6	3.6	-3.96	-8.68	15.68	75.34	34.37
4	7.5	6.7	-2.06	-5.58	4.24	31.14	11.49
5	9.1	8.9	-0.46	-3.38	0.21	11.42	1.55
6	11.2	6.6	1.64	-5.68	2.69	32.26	-9.32
7	13.4	14.3	3.84	2.02	14.75	4.08	7.76
8	13.7	24.5	4.14	12.22	17.14	149.33	50.59
9	14.2	20.8	4.64	8.52	21.53	72.59	39.53
10	16.2	30.5	6.64	18.22	44.09	331.97	120.98
合計	95.6	122.8	0	0	226.5	866.28	386.5
平均	9.56	12.28	-	-	$\uparrow S_{xx}$	$\uparrow S_{yy}$	$\uparrow S_{xy}$

グラフは下図です。



必要な各値を導出しましょう。

● $E[a] = a = \frac{S_{xy}}{S_{xx}} = \frac{386.5}{226.5} = 1.706$

● $V[a] = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{226.5}$

● $E[b] = b = -a\bar{x} + \bar{y} = -4.033$

● $V[b] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) = \sigma^2 \left(\frac{1}{10} + \frac{9.56^2}{226.5} \right)$

と代入したいのですが、

σ^2 はどこから求めるの？と不明ですよ！

実際は不明です。なので、**それらしい値**を使います。分散分析に慣れているとピンと来ますよね！

そうです！ V_e ですね！

●回帰分析では $V_e = \frac{S_e}{\varphi_e} = \frac{S_e}{n-2} = \sigma^2$ を使います。

● $S_e = S_T - S_R = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 206.76$ より、 $V_e = 206.76 / (10-2) = 25.85$ となります。

(3) グラフを描いてみる

例えば、 2σ 分ズレた場合の回帰直線を描きます。定数 a, b は以下のように変化します。

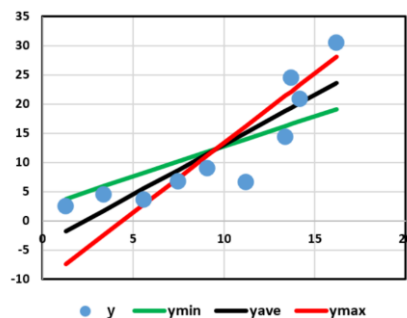
● $a_{max,min} = E[a] \pm m \times \sqrt{V[a]}$

● $b_{max,min} = E[b] \pm m \times \sqrt{V[b]}$ (ここで 2σ 分を考えるので、 $m=2$ とします)。

値は下表のようになります。

-	σ	a	b	回帰直線
min	-2	1.03	2.43	$y = 1.03x + 2.43$
ave	0	1.71	-4.03	$y = 1.71x - 4.03$
max	2	2.38	-10.49	$y = 2.38x - 10.49$

グラフを描くと下図のようになります。



回帰直線の区間推定をグラフで図示すると

●領域がわかる

●必ず平均 $(\bar{x}, \bar{y})$ を通る

ことがわかりますね。

以上、難しい公式を導出して、回帰直線の区間推定領域を求める過程を解説しました。
難しい導出過程でしたが、ちゃんとできましたね！公式は導出できてから使いましょう。

以上、「回帰直線の推定区間が導出できる」を解説しました。

【1】符号検定と検定統計量の定義

(1) 符号検定と検定統計量とは

ノンパラメトリック検定の一種で、次のように定義します。よく「1 標本検定」とか言われます。

$x_1, \dots, x_n$ を同一母集団からの大きさ n の標本とし、その分布の中央値を θ とする。

このとき、「母集団分布の中央値は θ_0 である」という仮説を考える。

●帰無仮説: $H_0: \theta = \theta_0$

●対立仮説: $H_1: \theta \neq \theta_0$

この場合、検定統計量を以下とし、 $N(0,1^2)$ に従う正規分布とする。

● $Z = \frac{U - E[U]}{\sqrt{V[U]}} \approx N(0,1^2)$

ただし、統計量 U, Z_i は

● $U = \sum_{i=1}^n Z_i$

● $Z_i = \begin{cases} 1 & (X_i \geq \theta_0) \\ 0 & (X_i < \theta_0) \end{cases}$

つまり、中央値以下なら Z_i は 0 で、中央値以上なら 1 とします。

これは、 Z_i は中央値前後に同数個あるとして、

確率 $P=1/2$ を使って統計量の期待値と分散を計算するためです。

定義は難しいですが、統計量の期待値と分散の計算は簡単です。

【2】検定統計量の期待値と分散は 2 項分布で解ける

(1) 符号検定のポイントは、二項定理と確率 $P=1/2$ であること

大波の相関、小波の相関でも話しますが、ポイントは、

二項定理と確率 $P=1/2$ を使うこと

ここがわかれば、

別に確率 $P=1/2$ を使わなくてもいいじゃん！と気づくはず

これを知らないと、大波の相関と小波の相関しか知らないとなってしまいます。

なので、

● Z_i が中央値以下である確率 $\Pr\{Z_i\}=0$

● Z_i が中央値以上である確率 $\Pr\{Z_i\}=1$

を使うと、統計量 U が θ_i より大きくなる確率 P は二項定理から

$$P = \sum_{r=0}^i {}_nC_r \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{n-r}$$

と計算できます。よくみると、

確率 $P=1/2$ の二項分布の期待値と分散を計算すればよいってことです。

ここまで準備して、検定統計量の期待値と分散を計算します。

(2) 検定統計量の期待値を導出

二項分布の期待値は

$E[U] = np$ ですね。ここに $p=1/2$ を代入します。

$E[U] = n/2$

(3) 検定統計量の分散を導出

二項分布の分散は

$V[U] = np(1-p)$ ですね。ここに $p=1/2$ を代入します。

$V[U] = n/4$

【3】符号検定表は自分で作れる

(1) 符号検定表とは

符号検定では、比較したい2つの値の差から、符号検定表を使って検定します。

N	有意水準 $\alpha=1\%$	有意水準 $\alpha=5\%$
10	0	1
11	1	2
12	1	2
13	1	3
14	2	3
15	2	3
16	2	4
17	3	4
18	3	5
19	4	5
20	4	5
25	6	7
30	8	10
50	16	18
100	37	41

符号検定表は自分で作れる！自分で作ると符号検定の意味がよくわかる！

(2) 符号検定表の値の求め方を理解せよ

符号検定表の値はどうやって計算するかですが、

●数式で書くと

$$\Pr = \sum_{r=0}^i {}_n C_r p^r (1-p)^{n-r} \geq \alpha \text{ (有意水準)}$$

を満たす最小の*i*を考えると、

i - 1 は $\Pr \leq \alpha$ を満たす最大の整数になるので、

i - 1の値が符号検定表の値になる。

検定統計量の期待値と分散にも出てきた、二項分布の式がしっかり入っていますね。

●Excel の式で書くと

`BINOM.inv(個数 N, 確率 P, 有意水準  $\alpha$ )-1`

を代入します。

1つ例を挙げて計算しましょう。 $n=9$ $p=1/2$ で $\alpha=0.05$ の場合

$$\Pr(i=0) = {}_9 C_0 \left(\frac{1}{2}\right)^0 \left(1 - \frac{1}{2}\right)^{9-0} = 0.00195$$

$$\Pr(i=1) = {}_9 C_1 \left(\frac{1}{2}\right)^1 \left(1 - \frac{1}{2}\right)^{9-1} = 0.01953$$

$$\Pr(i=2) = {}_9 C_2 \left(\frac{1}{2}\right)^2 \left(1 - \frac{1}{2}\right)^{9-2} = 0.08984$$

...

と計算していきます。

ここで、

$\Pr(i=0) + \Pr(i=1) < \alpha = 0.05$ ですが、

$\Pr(i=0) + \Pr(i=1) + \Pr(i=2) > \alpha = 0.05$ ですね。

なので、*i*=2 となるので、符号検定表の値は *i*-1=1 となります。

これをコツコツ計算すれば OK ですが、計算より、方法を理解しましょう。二項分布の式を使っているだけです。

【4】大波の相関、小波の相関の本質を理解せよ。

では、符号検定から「大波の相関」、「小波の相関」があるわけですが、この分類はどこから来たのでしょうか？ これを理解するのは解き方を暗記するより、はるかに重要です！

符号検定表の値を求める式は、検定統計量の式とよく似た

$$\Pr=\sum_{r=0}^i {}_nC_r p^r (1-p)^{n-r} \geq \alpha \text{ (有意水準)}$$

ですよね。この式から、「大波の相関」、「小波の相関」を意味する変数があります。どれか？わかりますか？ ヒントは、すでに本記事で話したとおり、

大波の相関、小波の相関は、確率 $P=1/2$ である場合を意味する。

確かに、

●大波の相関：中央値と比較

●小波の相関：前値と比較

ですから、

●大波の相関→中央値と比較して大小どちらかに割れる確率は $1/2$ に近い

●小波の相関→前の値と比較して大小どちらかに割れる確率は $1/2$ に近い

という意味です。

(1) 基本は符号検定表の値を求める式は同じ

なぜなら、大波の相関も小波の相関も符号検定表を使うわけですから、符号検定表の値を求める式は

$$\Pr=\sum_{r=0}^i {}_nC_r p^r (1-p)^{n-r} \geq \alpha \text{ (有意水準)}$$
と同じです。

この式で表現できるケースがいくつかあり、その中の1つが大波の相関と小波の相関と言うわけです。

(2) 大波、小波は何を意味するのか？

特に意味はなく、確率 $p=1/2$ として計算できるケースのうち2つが大波の相関と小波の相関ということです。大波と小波で対義語が使われているので、2つしかないように思いますが、数式から考えるとそうではないことがわかりますね。

【5】大波、小波以外の相関も考えてみよう。

数式を再掲すると

$$\Pr=\sum_{r=0}^i {}_nC_r p^r (1-p)^{n-r} \geq \alpha \text{ (有意水準)}$$
ですから、別に $p=1/2$ 以外を使っても問題ありません。後は運用しやすいかどうかだけです。

是非 大波、小波の相関以外の検定も見つけてみてください。

以上、「大波の相関、小波の相関、符号検定がよくわかる」を解説しました。

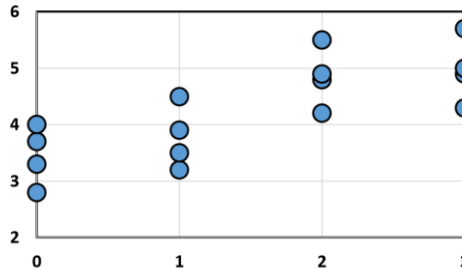
繰返しのある単回帰分析がわかる

- (i) 平方和、回帰の分散分析、回帰直線の求め方は、繰返し有り無し関係なく 1 つの解法で OK
(ii) 分散分析は実験計画法（一元配置実験）と比較しながら理解を深めよう！

【1】繰返しのある単回帰分析とは

(1) 繰返し有無の違い

繰返しのある単回帰分析とは、**同じ x で y の値が同じ複数個データがある特殊な場合**です。
グラフで繰返しの有りを理解しましょう。



(2) 繰返しがあると何に注意すべきか？

結論から言うと、

- (i) 平方和、回帰の分散分析、回帰直線の求め方は、繰返し有り無し関係なく 1 つの解法で OK
(ii) 分散分析は実験計画法（一元配置実験）と比較しながら理解を深めよう！

でしたね。これは、データの構造式で説明できます。後で解説します。

(3) 本記事で扱うデータを用意

次の 16 組のデータを用意します。グラフを描くと上図になります。

i	j	x_i	y_{ij}
1	1	0	3.3
1	2	0	2.8
1	3	0	3.7
1	4	0	4
2	1	1	4.5
2	2	1	3.9
2	3	1	3.2
2	4	1	3.5
3	1	2	4.8
3	2	2	4.2
3	3	2	5.5
3	4	2	4.9
4	1	3	4.9
4	2	3	5.7
4	3	3	5
4	4	3	4.3
-	合計	24	68.2
-	平均	1.5	4.2625

上表で、 $i(=1,2,3,4)$ が変数 x_i が変化する値で、 $i(=1,2,3,4)$ が同じ変数 x_i で繰り返す値となり、 $4 \times 4=16$ 個のデータとなっています。

【2】平方和 S_{xx} , S_{xy} , S_{yy} を計算

(1) 注意してほしい点

繰返し有無関係なく、同じ解き方で計算します。

i	j	x	y	A= (x - $\bar{x}$)	B= (y - $\bar{y}$)	A ²	B ²	AB= (x - $\bar{x}$)(y - $\bar{y}$)
1	1	0	3.3	-1.5	-0.96	2.25	0.93	1.44
1	2	0	2.8	-1.5	-1.46	2.25	2.14	2.19
1	3	0	3.7	-1.5	-0.56	2.25	0.32	0.84
1	4	0	4	-1.5	-0.26	2.25	0.07	0.39
2	1	1	4.5	-0.5	0.24	0.25	0.06	-0.12
2	2	1	3.9	-0.5	-0.36	0.25	0.13	0.18
2	3	1	3.2	-0.5	-1.06	0.25	1.13	0.53
2	4	1	3.5	-0.5	-0.76	0.25	0.58	0.38
3	1	2	4.8	0.5	0.54	0.25	0.29	0.27
3	2	2	4.2	0.5	-0.06	0.25	0	-0.03
3	3	2	5.5	0.5	1.24	0.25	1.53	0.62
3	4	2	4.9	0.5	0.64	0.25	0.41	0.32
4	1	3	4.9	1.5	0.64	2.25	0.41	0.96
4	2	3	5.7	1.5	1.44	2.25	2.07	2.16
4	3	3	5	1.5	0.74	2.25	0.54	1.11
4	4	3	4.3	1.5	0.04	2.25	0	0.06
-	合計	24	68.2	0	0	20	10.6	11.3
-	平均	1.5	4.2625	-	-	$\uparrow$ S_{xx}	$\uparrow$ S_{yy}	$\uparrow$ S_{xy}

(2) S_{xx}, S_{xy}, S_{yy} を式で書く。

$$\bullet S_{xx} = \sum_{i=1}^a \sum_{j=1}^b (x_i - \bar{x})^2 = b \sum_{i=1}^a (x_i - \bar{x})^2$$

$$\bullet S_{xy} = \sum_{i=1}^a \sum_{j=1}^b (x_i - \bar{x})(y_i - \bar{y})$$

$$\bullet S_{yy} = \sum_{i=1}^a \sum_{j=1}^b (y_i - \bar{y})^2$$

と機械的に書けます。

【3】 回帰の分散分析を計算

(1) 基本は、繰返し有無に意識しなくていい

平方和 S_{xx}, S_{xy}, S_{yy} から回帰分析の分散分析と、回帰直線に必要な傾きと y 切片が計算できます。これも、繰返し有無関係なく 1 つの解法でできます。

(2) 平方和 S_T, S_R, S_{er} を計算

単回帰分析では、

$$\bullet \text{総平方和} : S_T = S_{yy} = 10.60$$

$$\bullet \text{回帰平方和} : S_R = \frac{S_{xy}^2}{S_{xx}} = 6.38$$

$$\bullet \text{回帰残差平方和} : S_{er} = S_T - S_R = 4.22$$

から計算できて、これは繰返し有無関係なく同じ解法ですね。

分散分析表を作ると

-	S	Φ
R	6.38	1
er	4.22	14
T	10.60	15

ですね。繰返し有無関係なく、回帰の分散分析はできますね。

これも同様に、繰返し有無関係なく1つの解法でできますね。

●傾き : $\beta_1 = \frac{s_{xy}}{s_{xx}} = 0.565$

$$y = 0.565x + 3.415$$

になります。

ここからが、繰返し有りの場合で意識しておくべき内容です。

実験計画法(一元配置実験)と比較しながら、まとめて理解した方が分かりやすい！

繰返し有りの回帰分析でも、実験計画法(一元配置実験)にもなっていますね。

<繰返し有りの回帰分析>

i	j	x_i	y_{ij}
1	1	0	3.3
1	2	0	2.8
1	3	0	3.7
1	4	0	4
2	1	1	4.5
2	2	1	3.9
2	3	1	3.2
2	4	1	3.5
3	1	2	4.8
3	2	2	4.2
3	3	2	5.5
3	4	2	4.9
4	1	3	4.9
4	2	3	5.7
4	3	3	5
4	4	3	4.3

<実験計画法(一元配置実験)>

A	y_{ij}			
A0	3.3	2.8	3.7	4
A1	4.5	3.9	3.2	3.5
A2	4.8	4.2	5.5	4.9
A3	4.9	5.7	5	4.3

下図を見るとよくわかります。

繰返し有りの単回帰分析: $(y_{ij} - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_{ij} - \hat{y}_i)$
 回帰R 残差er

一元配置実験: $(y_{ij} - \bar{y}) = (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i)$
 効果A 残差e

両者混ぜると: $(y_{ij} - \bar{y}) = (\hat{y}_i - \bar{y}) + (\bar{y}_i - \hat{y}_i) + (y_{ij} - \bar{y}_i)$
 回帰R lof 効果A 残差e

$S_T = S_R + S_{lof} + S_e$
 S_A S_{er}

以上、「繰返しのある単回帰分析がわかる」を解説しました。

【1】重回帰分析のテコ比がベース

(1) テコ比の関連記事

重回帰分析の記事になりますが、テコ比、ハット行列の導出も計算も理解できる関連記事をまとめております。確認ください。この関連記事をベースに本記事を作っています。

【関連記事】重回帰分析のテコ比がよくわかる

<https://qcplanets.com/method/multi-regression/leverage1/>

(2) ハット行列とテコ比

関連記事の結果をまとめます。この式をベースに本記事で解説します。

- 回帰直線： $(\hat{y} - \bar{y}) = \beta(x_i - \bar{x})$
- 行列表記
 - ・ $\hat{Y} : (\hat{y} - \bar{y})$
 - ・ $Y : (y_i - \bar{y})$
 - ・ $X : (x_i - \bar{x})$
- 回帰直線の行列表記： $\hat{Y} = HY$
- ハット行列： $H = X(X^T X)^{-1} X^T$
- テコ比： $h_{ii} = H$ の (i, i) 成分
- 回帰直線の傾き： $\beta = (X^T X)^{-1} X^T Y$

【2】単回帰分析のハット行列とテコ比を導出

(1) 単回帰分析データを用意

単回帰分析として説明変数 x_i と目的変数 y_i ($i = 1, 2, \dots, n$)を用意します。

目的変数の回帰成分を $\hat{y}_i$ として、

$$\bullet (\hat{y} - \bar{y}) = \beta(x_i - \bar{x})$$

という回帰式が成り立っているとしましょう。

データは $i = 1, 2, \dots, n$ と n 個あるので、行列表記します。回帰成分 $\hat{Y}$ 、目的変数 Y を以下のように定義します。

$$\bullet \hat{Y} = \begin{pmatrix} \hat{y}_1 - \bar{y} \\ \vdots \\ \hat{y}_n - \bar{y} \end{pmatrix}, \quad \bullet Y = \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}$$

また、説明変数の行列 X も定義します。

$$\bullet X = \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix}$$

ここで、

- 回帰直線の行列表記： $\hat{Y} = HY$
- ハット行列： $H = X(X^T X)^{-1} X^T$

が成り立っているので、単回帰分析の場合のハット行列を導出してみましょう。

(2) 単回帰分析のハット行列を導出

まず、 $X(X^T X)^{-1} X^T$ の中の $(X^T X)^{-1}$ を計算します。

$$\bullet X^T X = (x_1 - \bar{x} \quad \dots \quad x_n - \bar{x}) \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix} = (x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 = S_{xx} \text{ (平方和) と定数になりますね。}$$

よって、

$$(X^T X)^{-1} = \frac{1}{S_{xx}} \text{ となります。}$$

なので、ハット行列 H は、

$$H = X(X^T X)^{-1} X^T = \frac{1}{S_{xx}} X X^T = \frac{1}{S_{xx}} \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix} (x_1 - \bar{x} \quad \cdots \quad x_n - \bar{x}) = \frac{1}{S_{xx}} \begin{pmatrix} (x_1 - \bar{x})^2 & \cdots & (x_1 - \bar{x})(x_n - \bar{x}) \\ \vdots & \ddots & \vdots \\ (x_1 - \bar{x})(x_n - \bar{x}) & \cdots & (x_n - \bar{x})^2 \end{pmatrix}$$

と $n \times n$ の正方行列になります。

【3】単回帰分析のテコ比を計算

テコ比 h_{ii} はハット行列の対角成分なので、ハット行列をじつと眺めると $h_{ii} = \frac{(x_i - \bar{x})^2}{S_{xx}}$ とわかりますね。

テコ比: $h_{ii} = H$ の (i, i) 成分

● 回帰直線の傾き: $\beta = (X^T X)^{-1} X^T Y$

(1) 単回帰直線の傾き β を行列から計算

単回帰直線の傾き β は、

$$\beta = (X^T X)^{-1} X^T Y = \frac{1}{S_{xx}} X^T Y$$

$X^T Y$ は行列なので、具体的に書くと次式になります。

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} = \frac{1}{S_{xx}} (x_1 - \bar{x} \quad \cdots \quad x_n - \bar{x}) \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} = \frac{1}{S_{xx}} \{ (x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y}) \} = \frac{S_{xy}}{S_{xx}}$$

となります。

【4】単回帰分析のハット行列とテコ比を計算

(1) データを用意

下表にデータを用意します。ついでに回帰分析もやりますが、折角なので、解いてみてください。

	x_i	y_i	A= $x_i - \bar{x}$	B= $y_i - \bar{y}$	A ²	B ²	AB	$\hat{y}_i$	$\hat{y}_i - \bar{y}$
1	2	-4	-38	-76	1444	5776	2888	-5.529	-23.529
2	6	0	-34	-72	1156	5184	2448	6.235	-11.765
3	13	36	-27	-36	729	1296	972	26.824	8.824
4	19	40	-21	-32	441	1024	672	44.471	26.471
合計	40	72	-120	-216	3770	13280	6980	-	-
平均	10	18	-	-	$\uparrow S_{xx}$	$\uparrow S_{yy}$	$\uparrow S_{xy}$	-	-

なお、回帰直線については、

● 傾き $\beta_1 = 2.94$

● y切片 = -11.41

● 相関係数 $\rho = 0.954$

● $\hat{y}_i = -11.41 + 2.94 x_i$

(2) 単回帰分析のハット行列を計算

では、ハット行列を上表の値から求めましょう。ここで、以下の式を定義します。

● $X_{ij} = (x_i - \bar{x})(x_j - \bar{x})$ とします。

$X_{11} \sim X_{44}$ の各値は以下のとおりになります。

$$\begin{array}{cccc}
X_{11}=64 & X_{12}=32 & X_{13}=-24 & X_{14}=-72 \\
X_{12}=32 & X_{22}=16 & X_{23}=-12 & X_{24}=-36 \\
X_{13}=-24 & X_{23}=-12 & X_{33}=9 & X_{34}=27 \\
X_{14}=-72 & X_{24}=-36 & X_{34}=27 & X_{44}=81
\end{array}$$

ハット行列の各成分の値がわかり、さらに $S_{xx}=170$ で割れば、ハット行列は計算できます。よって、

$$H = \begin{pmatrix} 0.376 & 0.188 & -0.141 & -0.424 \\ 0.188 & 0.094 & -0.071 & -0.212 \\ -0.141 & -0.071 & 0.053 & 0.159 \\ -0.424 & -0.212 & 0.159 & 0.476 \end{pmatrix}$$

ハット行列が正しく計算できたかを検算するには、 $\hat{Y} = HY$ が一致するのを確認すればOKです。

$$HY = \begin{pmatrix} 0.376 & 0.188 & -0.141 & -0.424 \\ 0.188 & 0.094 & -0.071 & -0.212 \\ -0.141 & -0.071 & 0.053 & 0.159 \\ -0.424 & -0.212 & 0.159 & 0.476 \end{pmatrix} \begin{pmatrix} -22 \\ -18 \\ 18 \\ 22 \end{pmatrix} = \begin{pmatrix} -23.529 \\ -11.765 \\ 8.82 \\ 26.47 \end{pmatrix} = \hat{Y}$$

と確かに一致するので、計算が正しいことがわかります。

(3) 単回帰分析のテコ比を計算

$$\text{テコ比 } h_{ii} = \frac{(x_i - \bar{x})^2}{S_{xx}}$$

なので、 $i=1,2,3,4$ を代入すればOKです。

- $h_{11}=0.376$
- $h_{22}=0.094$
- $h_{33}=0.053$
- $h_{44}=0.476$

ところで、面白いことに、

$$\bullet h_{11} + h_{22} + h_{33} + h_{44} = 0.376 + 0.094 + 0.053 + 0.476 = 1$$

となることがわかっています。

つまり、**テコ比の合計は1になります！**
 なぜか？わかりますか？ 折角なので証明します。

(4) 単回帰分析のテコ比の合計が1になる理由

$$\text{テコ比 } h_{ii} = \frac{(x_i - \bar{x})^2}{S_{xx}}$$

より、

$$\sum_{i=1}^n h_{ii} = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{S_{xx}} = \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{S_{xx}}{S_{xx}} = 1$$

以上、単回帰分析のテコ比を解説しました。面白い問題でしたね。