２標本の平均の差の検定

消費電力が低くなるように電球に改良を加えた結果、改良前後の消費電力は次の通りであった。

改良前 (x)	:	7.8	7.5	7.7	8.2	8.0
改良後 (y)	:	6.9	6.4	7.2	7.1	

（単位 W）

消費電力の分布は同じ母分散を持つ正規分布であるとするとき、次の問いに答えよ。

1) 標本平均 $(\bar{x}, \bar{y})$・標本分散 (s_x^2, s_y^2) を求めよ。

$\bar{x} = 7.84,\ s_x^2 = 0.0730,\ \bar{y} = 6.90.\ s_y^2 = 0.12666\cdots$

2) 共通の母分散の推定量 s^2 を求めよ。

$$s^2 = \frac{1}{4+3}(4 \times 0.0730 + 4 \times 0.1267) = \frac{0.6721}{7} = 0.096014\cdots$$

3) 消費電力が低くなるといえるかどうか有意水準１％で検定せよ。
消費電力が低くなるかどうかなので、帰無仮説と対立仮説は下記のようになる。

帰無仮説：消費電力が高くなるか同じである　H_0: $\mu_x \leqq \mu_y$
対立仮説：消費電力が低くなる　H_1: $\mu_x > \mu_y$

実際の検定のときは H_0: $\mu_x - \mu_y = 0$ v.s. H_1: $\mu_x - \mu_y > 0$ として検定する。

帰無仮説（$\mu_x - \mu_y = 0$）が真のときの統計量の値を求めると下記のとおりである。

$$t = \frac{(7.84 - 6.90) - (0)}{\sqrt{0.09601} \times \sqrt{\frac{1}{5} + \frac{1}{4}}} = \frac{0.94}{0.2079} = 4.5214\cdots$$

（精密に計算した場合 $4.5225\cdots$）

t は自由度 $(5-1)+(4-1) = 7$ の t 分布に従うことを使って判定を行う。

この問題は対立仮説が「＞」の片側検定なので、棄却域はグラフの片側に作る。t が自由度 9 の t 分布に従うことから、片側１％（＝両側２％）の点を表から求めると 2.998 である。$t = 4.521 > 2.998$ なので、帰無仮説を棄却する。

つまり、改良後の方が消費電力が下がっていると言える。

注）今回の問題は**「消費電力が低くなる」**なので、対立仮説は $\mu_x > \mu_y$ の片側検定である。
「消費電力に違いがあるかどうか」の検定の場合、両側検定になる。
何を調べているのか問題文から読み取って正しい検定を行うこと。

確率統計　相関係数と回帰直線　　第１３回

・１変量から２変量（多変量）への拡張

今までは、身長なら身長だけ、体重なら体重だけのように、１つの変数に対して統計解析を行ってきた。しかしながら、身長と体重の間の関係を調べる場合、２つの変数を同時に扱う必要がある。そこで身長を x、体重を y で表し、(x, y) の組を１つの変量として扱うことを考える。n 人分のデータを得た場合、$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ となる。このとき、x_i と y_i は i 番目の人の（身長, 体重）の組であり、他人の身長・体重が組になることはない。

さらに（身長, 体重, 胸囲, 座高,···）のような多変数を組として１つの変量として扱うことも可能である。一般には p 変量を取り扱う場合、下記のようにベクトルと行列で表記するのが一般的であるが、ここでは２変量しか扱わないのでベクトル表記はしない。

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \quad \text{平均ベクトル :} \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \quad \text{分散共分散行列 :} \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}$$

μ_i : i 番目の変数の母平均（$E[X]$）
$\sigma_{ii} = \sigma_i^2$: i 番目の変数の母分散（$E[(X-\mu)^2]$）
σ_{ij} : i 番目の変数と j 番目の変数と共分散（$E[(X_i - \mu_{x_i})(X_j - \mu_{x_j})]$）

・相関係数とは

(x, y) の間に次のような関係があるとき

・x が増えると y も増える　　　　・x が増えると y が減る

その関係の強さをはかるものさしとして、相関係数 ρ（ギリシャ文字：ロー）が用いられる。

一般に n 個の標本 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ が得られたときの標本相関係数 r は

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1) \times s_x \times s_y}$$

のように、x, y の標本平均・標本分散と積和で表される。$-1 \leqq r \leqq 1$ の値を取り、絶対値が１に近いほど直線的な関係が見えるほか、

・r が正のとき：x が増えると y も増える。　散布図が右肩上がり
・r が負のとき：x が増えると y が減る。　散布図が右肩下がり

のような関係性がわかる。注意する点は相関係数は直線的な関係しかはかることができないので、２次関数上や円の上にある点のように、曲線的な関連性があっても相関係数の値は低くなる。

多変量の場合にも同様に定義することが可能で、i 番目の変数と j 番目の変数の相関係数 ρ_{ij} と分散 σ_i^2, σ_j^2、共分散 σ_{ij} の関係は次のとおりになる。

$$\sigma_{ij} = \rho_{ij}\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}} = \rho_{ij}\,\sigma_i\sigma_j$$

相関係数を並べた行列（形式上 $\rho_{ii} = 1$）を相関行列と呼ぶ。

・回帰直線について

回帰分析はデータの予測に用いられる手法で、いくつかの説明変数 $x_1, x_2, \cdots, x_p$ から目的変数 y を求める（予測する）関係式 $y = f(x_1, \cdots, x_p)$ を推定する。簡単な例として、おもりの重さとばねばかりの伸びは正比例の関係があることが知られている。つまりおもりの重さを x g、ばねばかりの伸びを y cm としたとき、

$$y = \alpha + \beta x$$

の関係がある（この直線を回帰直線と呼ぶ）。ばねばかりに関する情報が何も無い場合、何回かの実験によって得られた (x_i, y_i) の組で与えられる標本から α, β の値を推定する必要がある。測定にはある程度の誤差が考えられるので、すべての標本が $y = \alpha + \beta x$ 上に現れるわけではない。そこで次のような誤差 ε_i を考慮した関係式

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

を考えて、誤差を最小にする α, β の値を求める。誤差の単純和は正の値と負の値が打ち消しあってしまうため、実際には誤差の 2 乗和を最小にする α, β の値を計算する。詳しい説明は省くが、関数の最大値を求める問題なので、関数を α, β で偏微分し、0 となる α, β を求めればよい。実際に n 個の標本が与えられたとき、$\sum \varepsilon_i^2$ を最小にする α, β の推定値 $\hat{\alpha}, \hat{\beta}$ は次のようになる。

$$\begin{aligned}
\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \\
\hat{\beta} &= \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x^2}
\end{aligned}$$

つまり平均および分散と x, y の積和があれば求めることが可能である。ちなみに相関係数はこの回帰直線にどれだけ近いかを表す統計量と考えることも可能である。

例．次のデータの相関係数と回帰直線を求めよ。

番号	1	2	3	4	5	6
x	3	4	5	6	7	8
y	11.8	19.4	18.5	16.1	20.5	30.1

x と y のそれぞれの標本平均、標本分散を求めると

$\bar{x} = 5.5, \quad s_x^2 = 3.50 \qquad \bar{y} = 19.4, \quad s_y^2 = 37.032$

相関係数と回帰直線を求めるためには上記以外に $\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$ を求めればよい。

《$\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ の場合》

番号	1	2	3	4	5	6	計
x	3	4	5	6	7	8	33
y	11.8	19.4	18.5	16.1	20.5	30.1	116.4
$x-\bar{x}$	-2.5	-1.5	-0.5	0.5	1.5	2.5	0
$y-\bar{y}$	-7.6	0	-0.9	-3.3	1.1	10.7	0
積	19	0	0.45	-1.65	1.65	26.75	46.2

$$\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=46.2\right)$$

相関係数：$r=\dfrac{46.2}{(6-1)\sqrt{3.50}\sqrt{37.032}}=0.81161\cdots$

回帰直線：$\beta=\dfrac{46.2}{(6-1)\times 3.50}=2.64,\quad \alpha=19.4-2.64\times 5.5=4.88$

より回帰直線は $y=2.64x+4.88$

《$\displaystyle\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}$ の場合》

番号	1	2	3	4	5	6	計
x	3	4	5	6	7	8	33
y	11.8	19.4	18.5	16.1	20.5	30.1	116.4
積	35.4	77.6	92.5	96.6	143.5	240.8	686.4

$$\left(\sum_{i=1}^{n}x_iy_i=686.4\right)$$

相関係数：$r=\dfrac{686.4-6\times 5.5\times 19.4}{(6-1)\sqrt{3.50}\sqrt{37.032}}=\dfrac{46.2}{(6-1)\sqrt{3.50}\sqrt{37.032}}=0.81161\cdots$

回帰直線：$\beta=\dfrac{686.4-6\times 5.5\times 19.4}{(6-1)\times 3.50}=2.64,\quad \alpha=19.4-2.64\times 5.5=4.88$

より回帰直線は $y=2.64x+4.88$

関数電卓の種類によってはデータを入力するだけで相関係数及び回帰直線を求めることが可能なものもある（２変数統計機能)。テストの際は使い方を自分で調べて途中計算無しに関数電卓の結果を使ってもかまわない。

（ただし途中計算が無い場合、部分点が付かない〔つまり０点か満点か〕）

資料置場

https://www.gen.kanagawa-it.ac.jp/takeda/class/

2019年度 確率統計S　小テスト解答用紙　2020. 1.10

・回帰直線と相関係数

次のデータが与えられたときの、回帰直線と相関係数を求めよ。

番号	1	2	3	4	5	6
x	7	9	11	13	15	17
y	62.1	66.6	50.1	46.4	41.3	35.3

2019年度神奈川工科大学 確率統計S 演習問題	学科	学年	組	学　籍　番　号	氏　　　名	

提出先：K3-3309号室前　18番のボックス　　提出期限：　1月15日（水）昼休みまで