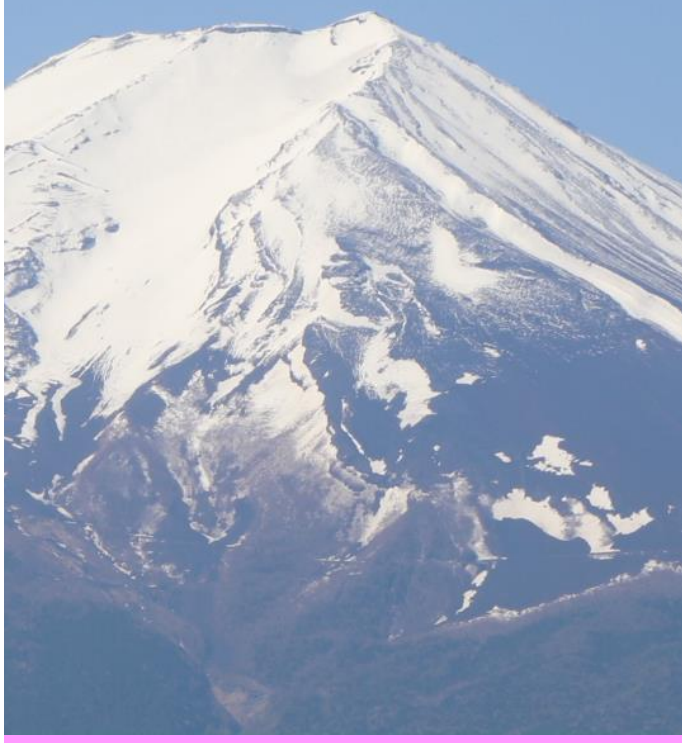
富士未来学Ⅴ

－統計分析講座



統計分析講座でできるようになること

２変量の関係に着目し、相関関係を調べたり、回帰分析を行ったりすることができる。統計に関する諸定理や法則を学び、推測統計の考え方を理解し、無相関検定や対応のあるｔ検定、適合度の検定、独立性の検定を行うことができる。

統計分析講座で学ぶこと

具体的な演習をとおして、回帰分析や無相関検定や対応のあるｔ検定、適合度の検定、独立性の検定の手法を学ぶ。

月　　日（　　）

東京都立富士高等学校

東京都立富士高等学校附属中学校

SSH Super Science Highschool

高校２年　　組　　番　氏名

統計分析講座①

ルーブリックを見返しながら、自己の到達度を確認し、改善を図りましょう。

育成したい資質・能力	グランドデザインの観点	評価の観点	評価の対象	高度に達成されている A	達成されている B	一部に課題あり C	自己評価	教員による評価
理数的解決力	分析解析	思考・判断・表現	4と6と7の記述	４と６と７において、全ての問題に自力で取り組み、答え合わせをし、間違えた問題については解説を参考にしながら直している。	４と６と７において、答え合わせをし、間違えた問題については解説を参考にしながら直しているが、自力で取り組んでいない問題がある。	４と６と７において、答え合わせをしていなかったり、解説を参考にしながら直していなかったりしている。		
理数的解決力	分析解析	思考・判断・表現	8と10の記述	８と１０において、全ての問題に自力で取り組み、答え合わせをし、間違えた問題については解説を参考にしながら直している。	８と１０において、答え合わせをし、間違えた問題については解説を参考にしながら直しているが、自力で取り組んでいない問題がある。	８と１０において、答え合わせをしていなかったり、解説を参考にしながら直していなかったりしている。		

1　回帰分析とは

変量 x と変量 y の間に、y が x で説明できるような関係があるとき、x を独立変数（説明変数）、y を従属変数（目的変数）といいます。例えば、重りの重さ x〔ｇ〕とばねの伸び y〔ｃｍ〕、年齢 x〔歳〕と血圧 ｙ〔mmHg〕などがあります。

図１のように、変量Ｘと変量Ｙには正の相関関係があり、およそ１次式（ $y = ax + b$ ）の関係があることが読み取れます。この直線を回帰直線といいます。データに回帰直線をあてはめて解釈することを、回帰分析といいます。特に、独立変数が一つであるとき、単回帰分析といいます。

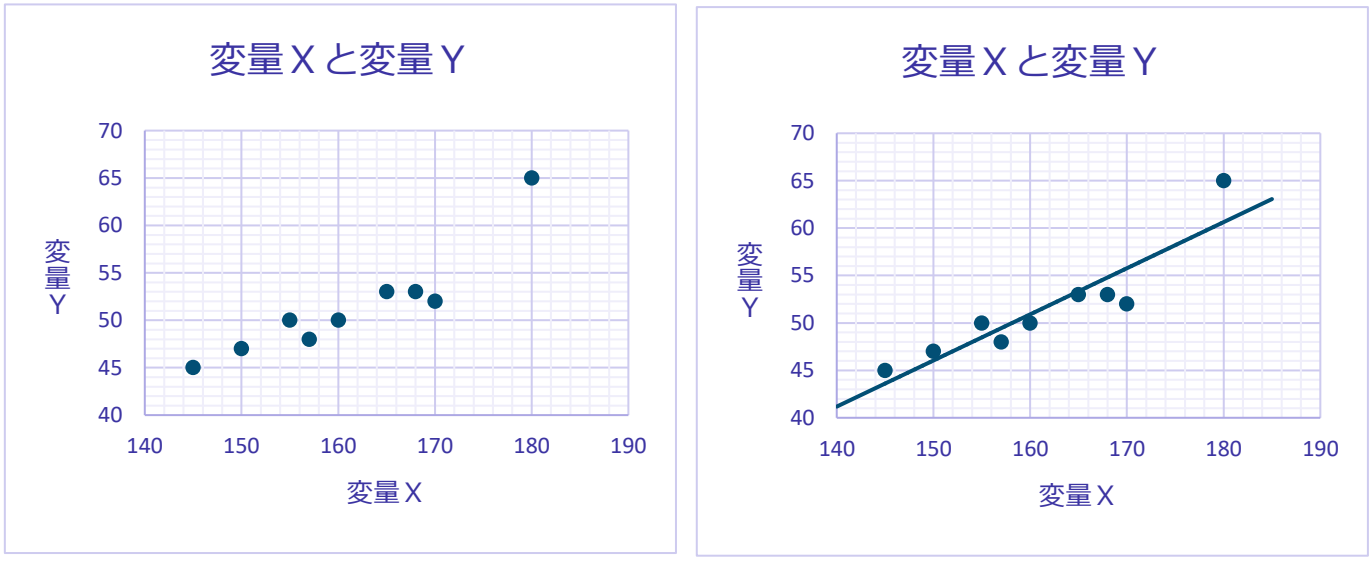


図１　１次式の関係

2　回帰と回帰直線

図１の右図のように、いくつかの点の配列を１本の曲線で代表することを回帰といいます。特に、１本の直線で回帰するとき、その直線 $y = ax + b$ を y の x への回帰直線といいます。また、傾きと切片を回帰係数といいます。

回帰直線は、それぞれの点の近くを通るようにしなければなりません。それぞれの点の近くを通るような回帰直線はどのように求めたらよいでしょうか。

3　回帰直線の公式

変量 x と変量 y の平均をそれぞれ $\overline{x}$ と $\overline{y}$ 、標準偏差をそれぞれ S_x と S_y 、共分散を S_{xy} 、相関係数を r とします。

直線 $y = ax + b$ を y の x への回帰直線とするとき、

$$a = \frac{S_{xy}}{S_x^{\,2}} = \frac{(x\text{ と }y\text{ の共分散})}{(x\text{ の分散})}\ 、\ b = \overline{y} - a\overline{x} \qquad ※\ a = r\frac{S_y}{S_x}$$

証明は、大学数学の内容を利用しますので、考え方の確認に留めます。図２を見てみましょう。各点の y 座標 y_i と直線上の値 $\hat{y}_i = ax_i + b$ の差（残差といいます）の２乗の値の和が最小になるように考えます。この方法を、最小二乗法といいます。

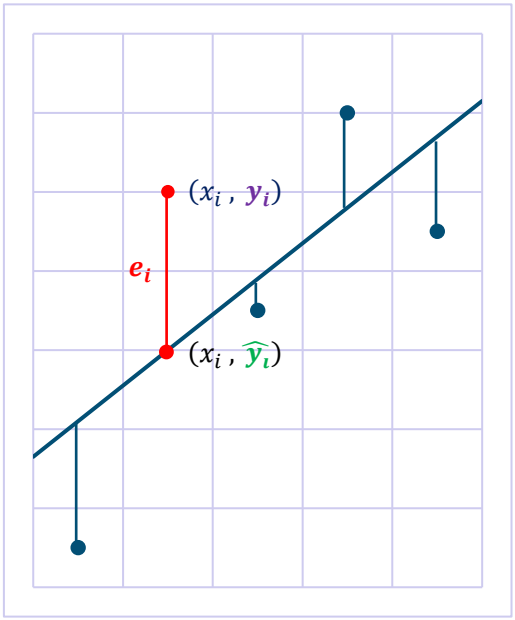


図２　最小二乗法

実現値（実際のデータ）である「 y_i 」と、予測値（直線上のデータ）である「 $\hat{y}_i$ 」の差が残差「 e_i 」です。最小二乗法では、残差の２乗の和が最小になるように考えます。

最小二乗法には偏微分などの計算が必要になります。この講座では、証明は省略し、結果を利用することにします。

4　回帰直線を求めましょう

次の表１は、ばねの変位を測定して得られたデータです。

表１　ばねの変位

	1	2	3	4	5
荷重 x	0	1	2	3	4
伸び y	１０	１４	２１	２６	２９

①荷重、伸びのそれぞれの平均値を求めましょう。

荷重の平均値 $\overline{x}$　　伸びの平均値 $\overline{y}$

②荷重の分散 ${S_x}^2$ を求めましょう。

$$\boldsymbol{{S_x}^2 = \frac{1}{n}\{(x_1-\overline{x})^2 + (x_2-\overline{x})^2 + \cdots + (x_n-\overline{x})^2\}}$$

③荷重の標準偏差 S_x を求めましょう。

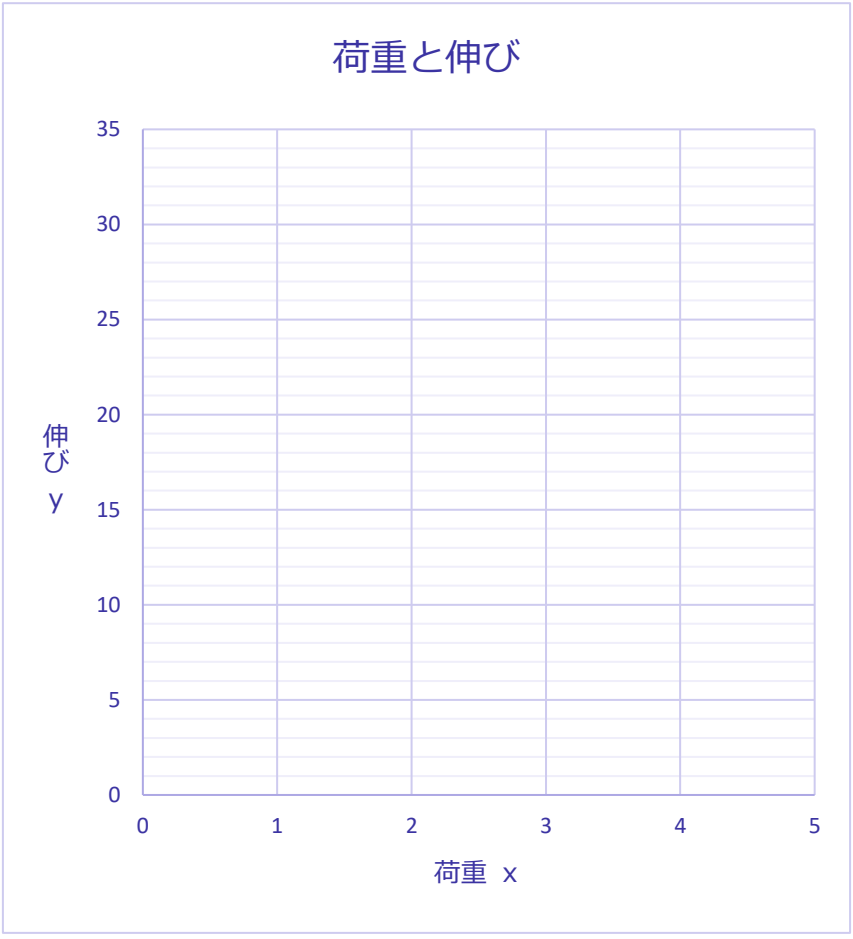
$$\boldsymbol{S_x = \sqrt{\frac{1}{n}\{(x_1-\overline{x})^2 + (x_2-\overline{x})^2 + \cdots + (x_n-\overline{x})^2\}}}$$

④荷重と伸びの共分散 S_{xy} を求めましょう。

$$\boldsymbol{S_{xy} = \frac{1}{n}\{(x_1-\overline{x})(y_1-\overline{y}) + (x_2-\overline{x})(y_2-\overline{y}) + \cdots + (x_n-\overline{x})(y_n-\overline{y})\}}$$

⑤回帰直線の傾き a と切片 b を求め、回帰直線の方程式を求めましょう。

⑥散布図を作成し、回帰直線のグラフもかきましょう。



⑦求めた回帰直線の方程式を用いて、荷重 $x = 2.5$ に対する伸び y を求めましょう。

5　決定係数とは

決定係数は、求めた回帰直線のあてはまりのよさを表します。求めた回帰直線が「どのくらいあてはまっているか」は、「従属変数が独立変数でどれだけ説明されるか」で考えます。そして、「従属変数が独立変数でどれだけ説明されるか」は「従属変数の分散が独立変数の分散でどれだけ説明されるか」で考えます。決定係数は、従属変数の値を独立変数がどれだけ説明しているかを表しています。課題研究で回帰分析を用いるときは、決定係数を記載しましょう。

回帰直線が「どのくらいあてはまっているか」

「従属変数が独立変数でどれだけ説明されるか」

「従属変数の分散が独立変数の分散でどれだけ説明されるか」

図３　決定係数（分散説明率）

決定係数には、いくつかの定義があります。本講座では、分散説明率を基に決定係数を説明します。まずは合成変数の平均と分散を考えます。

6　合成変数の平均

変量 x_n と変量 y_n の平均を、それぞれ $\overline{x}$ と $\overline{y}$ とします。それぞれのデータの数を n とします。このとき、変量 x_n と変量 y_n の合成変数 $v_n = cx_n + dy_n$ の平均について、次の式を証明しましょう。

$$\overline{v} = c\overline{x} + d\overline{y}$$

変量 x_n と変量 y_n の合成変数 $v_n = cx_n + dy_n$ の平均について、 $\overline{v} = c\overline{x} + d\overline{y}$ となることを証明しましょう。

<自分で解くためのスペース>

$$\overline{v} = \frac{1}{n}\sum_{i=1}^{n} v_i \ = \frac{1}{n}\sum_{i=1}^{n}(cx_i + dy_i) =$$

<解説を書くためのスペース>

※ $E(cx + dy) = cE(x) + dE(y)$、$E(cx + d) = cE(x) + d$ と表すこともあります。

7　合成変数の分散

変量 x_n と変量 y_n の平均を、それぞれ $\bar{x}$ と $\bar{y}$ とします。それぞれのデータの数を n とします。このとき、変量 x_n と変量 y_n の合成変数 $v_n = cx_n + dy_n$ の分散について、次の式を証明しましょう。

$$\boldsymbol{{S_{cx+dy}}^2 = c^2{S_x}^2 + 2cdS_{xy} + d^2{S_y}^2}$$

変量 x_n と変量 y_n の合成変数 $v_n = cx_n + dy_n$ の分散について、

${S_{cx+dy}}^2 = c^2{S_x}^2 + 2cdS_{xy} + d^2{S_y}^2$ となることを証明しましょう。

＜自分で解くためのスペース＞

$${S_{cx+dy}}^2 = \frac{1}{n}\sum_{i=1}^{n}\{(cx_i + dy_i) - (c\bar{x} + d\bar{y})\}^2$$

＜解説を書くためのスペース＞

※ $V(cx+dy) = c^2V(x) + 2cdCov(x,y) + d^2V(y)$、$V(cx+d) = c^2V(x)$ と表すこともあります。

8　残差の性質

残差の性質を確認します。残差には、平均が０、独立変数と無相関、従属変数と無相関などの性質があります。

（１）残差の平均は０

①独立変数 x に基づく従属変数 y の平均と予測値 $\hat{y}$ の平均が一致することを確かめましょう。回帰直線の公式から、$\hat{y} = ax + \overline{y} - a\overline{x}$ が成り立っているとします。

＜自分で解くためのスペース＞

$\hat{y} =$

＜解説を書くためのスペース＞

②残差 e の平均が 0 になることを確かめましょう。$e = y - \hat{y}$ であるとします。

＜自分で解くためのスペース＞

$e =$

＜解説を書くためのスペース＞

（２）独立変数と残差は無相関

無相関を示すには、共分散が０であることがいえればよいです。独立変数 x と残差 e の共分散が０であることを確かめましょう。回帰直線の公式から、$\hat{y} = ax + \overline{y} - a\overline{x}$ 、$a = \frac{S_{xy}}{{S_x}^2}$ が成り立っているとします。

＜自分で解くためのスペース＞

$$S_{xe} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(e_i - \overline{e})$$

＜解説を書くためのスペース＞

（3）予測値と残差は無相関

無相関を示すには、共分散が 0 であることがいえればよいです。予測値 $\hat{y}$ と残差 e の共分散が 0 であることを確かめましょう。回帰直線の公式から、$\hat{y} = ax + \overline{y} - a\overline{x}$ 、$a = \frac{S_{xy}}{{S_x}^2}$ が成り立っているとします。

<自分で解くためのスペース>

$$S_{\hat{y}e} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - \overline{\hat{y}}\right)(e_i - \overline{e})$$

<解説を書くためのスペース>

9　分散説明率

７で証明したことから、次の式が導かれます。

$$S_{x+y}{}^2 = S_x{}^2 + 2S_{xy} + S_y{}^2$$

ここで、$e = y - \hat{y}$ から $y = \hat{y} + e$ です。

$$S_y{}^2 = S_{\hat{y}+e}{}^2 = S_{\hat{y}}{}^2 + 2S_{\hat{y}e} + S_e{}^2$$

また、８（３）で証明したことから $S_{\hat{y}e} = 0$ です。

$$\boldsymbol{S_y{}^2 = S_{\hat{y}}{}^2 + S_e{}^2 \cdots ①}$$

このように、従属変数の分散は、予測値の分散と残差の分散の和になります。さらに、次の式が成り立ちます。

$$S_{\hat{y}}{}^2 = a^2 S_x{}^2 = \left(r\frac{S_y}{S_x}\right)^2 S_x{}^2 = r^2 S_y{}^2 \cdots ②$$

①と②より、次の式が成り立ちます。

$$S_e{}^2 = (1 - r^2) S_y{}^2 \cdots ③$$

②と③から①の右辺を書き換えることができます。

$$\boldsymbol{S_y{}^2 = r^2 S_y{}^2 + (1 - r^2) S_y{}^2 \cdots ④}$$

①から④の結果から分かることは、従属変数の分散が、$r^2 : (1 - r^2)$ の割合で予測値の分散と残差の分散に分けられるということです。

$$\boldsymbol{S_y{}^2 = S_{\hat{y}}{}^2 + S_e{}^2 \quad \cdots ①}$$

$$\boldsymbol{S_y{}^2 = r^2 S_y{}^2 + (1 - r^2) S_y{}^2 \cdots ④}$$

予測値は、独立変数で説明できる部分です。r^2 の値が大きいほど、従属変数のうち独立変数で説明できる割合が高いといえます。これらのことから、相関係数の２乗「 r^2 」を分散説明率と言います。分散説明率は、従属変数の値を独立変数がどれだけ決定しているかを表しているので、決定係数（R^2で表す）ともいいます。分散説明率が１に近いほど、回帰直線があてはまっていると考えることができます。これらのことから、最小二乗法による単回帰分析における決定係数は、相関係数の２乗「 r^2 」と一致することが分かります。

１０　決定係数を求めましょう

９より、$R^2 = r^2$であることが分かりました。

４の表１のデータから決定係数を求めましょう。小数第５位を四捨五入しましょう。

<自分で解くためのスペース>

<解説を書くためのスペース>

１１　ルーブリックによる自己評価

育成したい資質・能力	グランドデザインの観点	評価の観点	評価の対象	高度に達成されている A	達成されている B	一部に課題あり C	自己評価	教員による評価
理数的解決力	分析解析	思考・判断・表現	４と６と７の記述	４と６と７において、全ての問題に自力で取り組み、答え合わせをし、間違えた問題については解説を参考にしながら直している。	４と６と７において、答え合わせをし、間違えた問題については解説を参考にしながら直しているが、自力で取り組んでいない問題がある。	４と６と７において、答え合わせをしていなかったり、解説を参考にしながら直していなかったりしている。		
理数的解決力	分析解析	思考・判断・表現	８と１０の記述	８と１０において、全ての問題に自力で取り組み、答え合わせをし、間違えた問題については解説を参考にしながら直している。	８と１０において、答え合わせをし、間違えた問題については解説を参考にしながら直しているが、自力で取り組んでいない問題がある。	８と１０において、答え合わせをしていなかったり、解説を参考にしながら直していなかったりしている。		

引用文献

(1) 南風原朝和（2021）『心理統計学の基礎―統合的理解のために』，有斐閣アルマ，pp.57-62.

(2) 大村平（2018）「相関係数はこれだ」「直線で回帰する」『多変量解析のはなし【改訂版】―複雑さから本質を探る―』，日科技連，pp.43-45，pp.75-79.

(3) 東京都立富士高等学校・東京都立富士高等学校附属中学校（2021）『令和３年度スーパーサイエンスハイスクール研究開発実施計画書【開発型・実践型】』