もう忘れない. データの分析.

kakekakemiya

1　はじめに

単元別数学攻略シリーズ, 次はデータの分析です. 定期テストや普段の演習では殆ど扱わないくせに, 共通テスト・マーク模試で重要項目として出てくる**高校数学随一の不人気さ**を誇る単元です. [*1]

高校数学のくせに暗記を要求され, 計算もだるいという悪しき単元ですが, 現代における重要性は割と高い[*2]上に, 共通テストでの得点源としやすいので, 是非できるようになりましょう.

このプリントは, 暗記大嫌いな宮崎が, 「覚えやすく, 忘れにくい各テーマの暗記法」を紹介するというものにしたいと思います. 少しはデータの分析が楽しく感じられるようになるかも？

※本プリントでは, 教科書では $\cdots$ を使って表されることの多い部分を $\sum$ を用いて表記しています. こちらの方が覚えやすく, 理解しやすいためですが, $\sum$ を習っていない方はご注意ください.

2　データを代表する値

まず, 平均値, 中央値, 最頻値という, 三つの値について話していきます. この三つは, どれも, あるデータ群がどれくらいの値を持っているかをざっとつかむための**代表値**と呼ばれるものです. まずはそれぞれの定義を見ていきましょう.

定義

それぞれの定義を見ていきます.

平均値 所謂平均です. $\dfrac{1}{n}\displaystyle\sum_{k=1}^{n} x_k$

中央値 データを大きい順（小さい順でも良い）に並べたとき, 真ん中に来るもの.
データ数が偶数の場合は, 真ん中二つの平均を取る. （← 重要）

最頻値 漢字の通り, 最も多く出現した値. $\{1, 2, 3, 4, 4, 5\}$ なら 4

各々の違いについて

突然ですが, 皆さんに質問です. この 3 つの代表値のうち, 優れているものはどれでしょう？ 1, 2, 3 と順位をつけてください.

おそらく, 「平均値が一番優秀で, 次に中央値, 一番使いにくいのが最頻値」と考えたのではないでしょうか？「これがメンタリズムです. [*3]」というのも, 初めてこの 3 つを習ったとき, 平均が一番優秀やろ！と思うのはある

[*1] 僕はこの単元が大好きという人に出会ったことがありません.
[*2] 統計学の基礎につながるためです.
[*3] 違います.

種必然的で, その上平均だけは小さい頃から使い慣れているというのもこの傾向に拍車をかけています. ただ, **平均より他の代表値を使った方がいい時もあるんだよ**ってのを今日は学んでください.

手始めに, 平均の弱点を見ていきましょう. たとえば, 100 人がゲームをやって, 一人だけが, 10000 点を, 他の 99 人が 0 点を取ったとしましょう. すると, 平均は $\frac{10000}{100} = 100$ 点 となります.

これは代表の値としてふさわしいでしょうか？

もっと極端な例を考えてみましょう. 一回 10 円の宝くじで, 100 枚から 1 枚だけ引けるとします. このとき,

- くじの全ての当選額が 100 円
- くじのうち 1 枚が 10000 円で他は全て 0 円

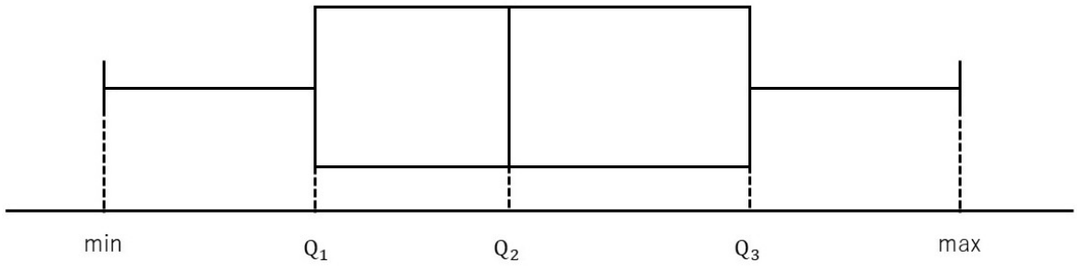
という 2 つの場合では, 同じ平均 100 円でも全くもって話が変わってきますよね？前者なら絶対に儲けられますが, 後者のくじはあまり買いたくありません. このように, 平均は, 極端に偏ったデータ群を扱うときにはあまり向いていないのです.

一方で中央値は, こういった偏ったデータに強く, 上の例でもしっかりと 0 を代表させます. 最頻値は, 少々毛色が違うというのが正直なところで, データがある程度丸められている（小数よりも整数, 小数点以下 3 桁よりも 2 桁）方が威力を発揮します.

この例だけで代表値の特色を理解できたとするのはあまりにも乱暴ですが, とりあえずは平均値が完璧でないことをおさえておいてくれれば良いです. [*4]

3　四分位範囲と箱ひげ図

四分位数は, 第 1 から第 3 まであります.



このときの, $Q_3 - Q_1$ を四分位範囲, その半分 $\dfrac{Q_3 - Q_1}{2}$ を四分位偏差といいます.

なんとなく四分位数, 箱ひげ図はいけてしまうという人が多いと思うので, ここでは豆知識を書いておくのにとどめます. [*5]

四分位数についての豆知識

以下の４つのパターンについて考えてみましょう.

1. $\{1,2,3,4,5\}$ のとき, $Q_1 = 1.5, Q_2 = 3, Q_3 = 4.5$

[*4] 他の特徴は各自考えてみてください.

[*5] 逆に苦手な人はしっかり確認しておくこと！

2. $\{1,2,3,4,5,6\}$ のとき, $Q_1 = 2, Q_2 = 3.5, Q_3 = 6$
3. $\{1,2,3,4,5,6,7\}$ のとき, $Q_1 = 2, Q_2 = 4, Q_3 = 6$
4. $\{1,2,3,4,5,6,7,8\}$ のとき, $Q_1 = 2.5, Q_2 = 4.5, Q_3 = 6.5$

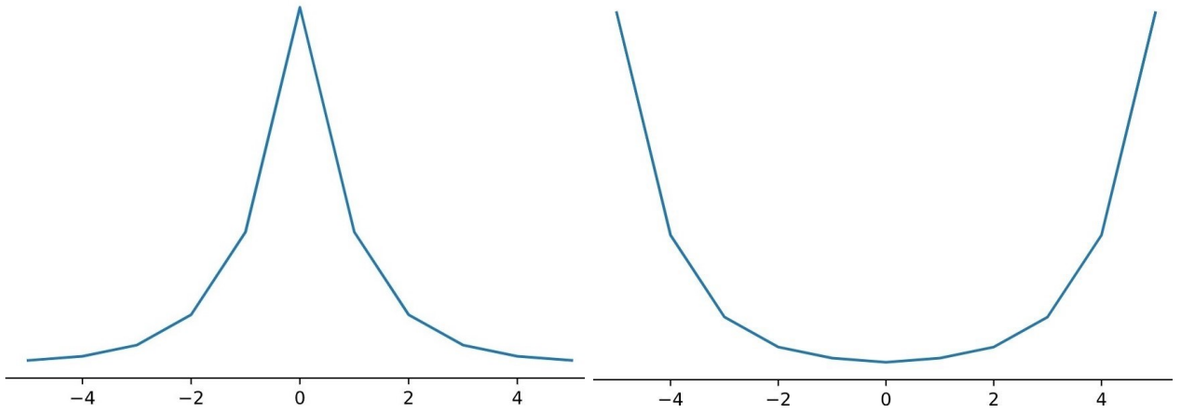
変な計算なく簡単に Q_i が求まるのは 3 番ですね！これはなぜでしょうか？考えればすぐに気づくと思いますが, データの個数を 4 で割ったときのあまりで分類されます.

すなわち, 4 で割ったとき, 3 余るものが一番楽ということになります. そして, **47 を 4 で割ったあまりは 3 です.** 何のこっちゃって感じかもしれませんが, 共通テストのデータのサンプルとして, 都道府県がよく出るんですね. これは統計を持ってきやすいのはもちろん, 四分位数が扱いやすいのも少なからず影響しているわけです. ということで, **47 個のデータの四分位数は何番目のデータか覚えておいてもいいんじゃない？**というのが豆知識でした.

4　分散と標準偏差

来ました, この辺りからアレルギーを発症する人が多いのではないでしょうか？よくわからんし, 全然楽しくない分野ですよね. ただ, しっかり理解すると, 割といろいろ考えられてるなあとわかると思うので, 是非読んでください.

そもそも分散って？



この二つのデータを比べてみてください. 横軸がスコア, 縦軸がそこに該当する人数です. この二つ, どちらも平均, 中央値は 0 点ですが, 明らかに特色が違いますよね.

左のグラフは 0 点付近の人が多く, 右のグラフは激しい二極化状態です. この二つを同じように扱うのはさすがに横着ですよね.

すなわち, 同じような代表値を持つ場合でも, **データのばらつき具合が違う**ことがあるのです. この, **ばらつき具合を表現するための指標が分散です.**

分散を定義する

それでは, どうしたらこのばらつき具合をうまく表現できるでしょうか？

各データが, 平均から離れているほどばらつき具合が大きいわけですから, **各データから平均を引いてあげて, (これを偏差といいます.) その合計を考えてみる**のはどうでしょう？割とシンプルな発想ですね. 式にすると,

分散の定義（仮）

$$\sum_{k=1}^{n}(x_k-\overline{x})$$

です.

なんだ, とっても簡単じゃないか！って感じですが, 実はこれ, うまくいきません. 例えば {1,2,3,4,5}(平均 3) では,

$$(1-3)+(2-3)+(3-3)+(4-3)+(5-3)=-2-1+0+1+2=0$$

となるんですが, 実はこれはどんなデータでも 0 になります. [*6]

というのも,

$$\begin{aligned}\sum_{k=1}^{n}(x_k-\overline{x}) &= \sum_{k=1}^{n}x_k-\sum_{k=1}^{n}\overline{x}\\ &= \sum_{k=1}^{n}x_k-\overline{x}\sum_{k=1}^{n}1\\ &= \sum_{k=1}^{n}x_k-n\overline{x}\\ &= \sum_{k=1}^{n}x_k-n\cdot\frac{1}{n}\sum_{k=1}^{n}x_k\\ &= \sum_{k=1}^{n}x_k-\sum_{k=1}^{n}x_k\\ &= 0\end{aligned}$$

となるからです. 平均はもともと全部の和を n で割ったものですから, 総和からそれを n 回引いてしまっては 0 になるのは言われてみれば当たり前ですね.

はてさて, 困ったものですね, 偏差の大きさがどれくらいか知りたいのに, 偏差の和を取ると打ち消して 0 になってしまうのです.

そこで考えられた解決策が, 偏差を二乗してから和を取る方法です. こうすれば, 各々は必ず正になるので, 打ち消してしまうことを避けられます.

分散の定義（仮）

$$\sum_{k=1}^{n}(x_k-\overline{x})^2\quad(\geq 0)$$

だんだんと分散っぽい形になってきましたね. ですが, これも少し問題があります.

この定義だと, {-10,0,10} の分散は平均が 0 だから $10^2+0+10^2=200$ となる一方で, {-1,-1,-1,$\cdots$,1,1,1} (-1,1 がそれぞれ 1000 個ずつ. 平均は 0) の分散は, $(-1)^2\times 1000+1^2\times 1000=2000$ となります.

ばらつき具合という点では明らかに前者の方が大きいのに, **要素数が多いというだけで**後者の分散の方が大きくなっているのです. となるとすなわち, **要素数に左右されずにばらつき具合を表現する必要があります.**

これを解決するためには, 皆さん大好き平均を使えばよいです. 「偏差の二乗の和」ではなく, 「偏差の二乗の和の平均」を考えるのです. 以上を踏まえると,

[*6] なんだと！？証明したいぜ！って方は自分なりに考えてから進んでください.

分散の定義

$$(\text{分散 } s_x^2) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2$$

となります.

今までよくわからない定義だなあと思っていた人も, 少しは親しみやすい形に見えてきたのではないでしょうか？

標準偏差について

ここまでで分散は何となくわかったけど, 標準偏差ってなんでいるん？そんな風に思っている方も多いのではないでしょうか. 先に定義を言ってしまうと, **標準偏差は分散の平方根**です.

標準偏差の定義

$$(\text{標準偏差 } s_x) = \sqrt{(\text{分散 } s_x^2)} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2}$$

なんかしょぼいというか, 存在価値…ってなりますよね. でも実は標準偏差はとても大事なのです.

{-2,-1,0,1,2} の分散は, $\frac{1}{5}\left\{(-2)^2 + (-1)^2 + 0 + 1^2 + 2^2\right\} = 2$ となりますが, この {-2,-1,0,1,2} というのが, A,B,C,D,E 君の前回のテストからの点数変動だとしましょう. すなわち, A 君は 2 点下がり, D 君は 1 点上がったというわけです.

さてこのとき, この分散 2 というのはどういう意味を持つでしょうか？ばらつきが 2 点くらいということでしょうか？いいえ違います. **ばらつき具合が, 2 点^2 という意味です.**

この見慣れない単位, 点^2 ってなんやねんって感じですね. もうお察しかと思いますが, **分散は単位がバグって扱いにくい**のです. よって, 単位を元に戻すために, $\sqrt{\ }$ をとってあげると扱いやすくなります. これこそが**標準偏差**です. [*7]

5 共分散と相関係数

さて, でました, 高校のデータのボスといっても過言ではないやつらです. 無理やりテスト前, 模試前に覚えて, 忘れてを繰り返している方が多いのではないでしょうか？[*8]

そもそも何がしたいのか？

これは即答できるようにしておいてくださいね. 共分散と相関係数はいずれも, **2 つの変量 x, y を持つデータについて, x, y の相関関係を捉える**のが目的です.

もっとシンプルに言うと, x が大きいほど y は大きくなるのか, 逆に小さくなるのか, はたまた 2 つは関係ないのか. というのを数字で表すのが目的です.

共分散

共分散を一から導くのは少々骨が折れるので, 先に定義をみて, その意味を考えていきましょう.

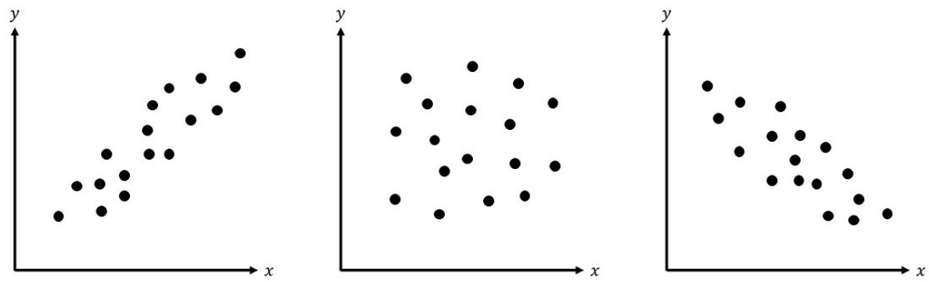
[*7] 科学実験などでは分散はあまり使われず, 標準偏差を使ってばかりだったりします.

[*8] 正味それでいい気もしますが笑

共分散の定義

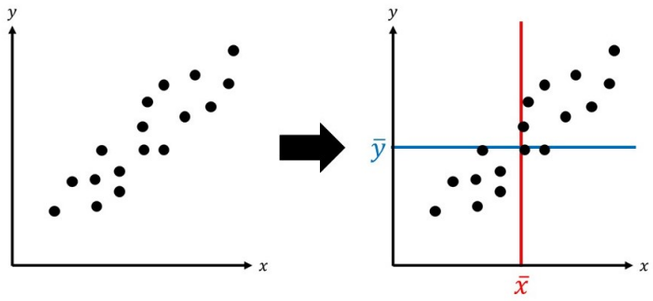
$$(\text{共分散 } s_{xy}) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y})$$

偏差の二乗の平均が分散であったのに対して, $(x_k - \overline{x})(y_k - \overline{y})$ の平均が共分散です. [*9]
さて, これが意味するところですが, 図を見ながら考えていきましょう.

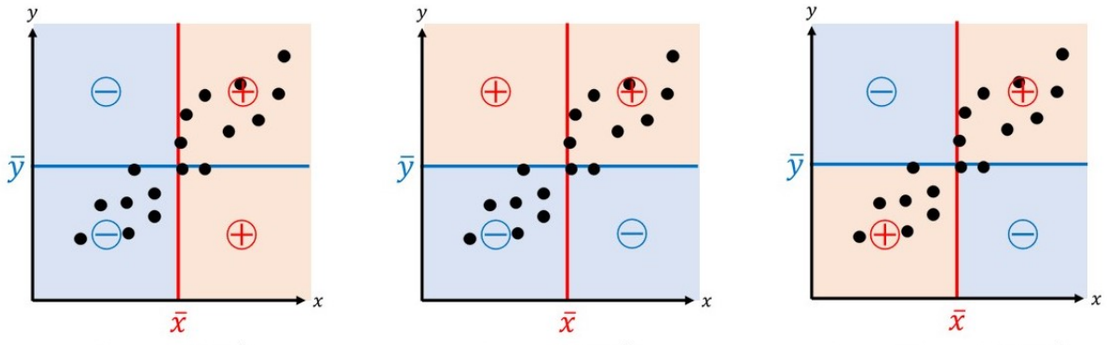


正の相関　　相関なし　　負の相関

まず, 平均 $\overline{x}, \overline{y}$ の位置を考えてみます. ここを基準にして, $(x_k - \overline{x})(y_k - \overline{y})$ が正か負かが変わってきます. 具



体的には, $(x_k - \overline{x})$ の正負, $(y_k - \overline{y})$ をそれぞれ考えてあげて, それをかけたものが $(x_k - \overline{x})(y_k - \overline{y})$ ことに注意すれば,



$(x_k - \bar{x})$の正負　　$(y_k - \bar{y})$の正負　　$(x_k - \bar{x})(y_k - \bar{y})$の正負

のようになります.

[*9] $y \to x$ とすれば, ただの分散の式になります. これでもう忘れない！

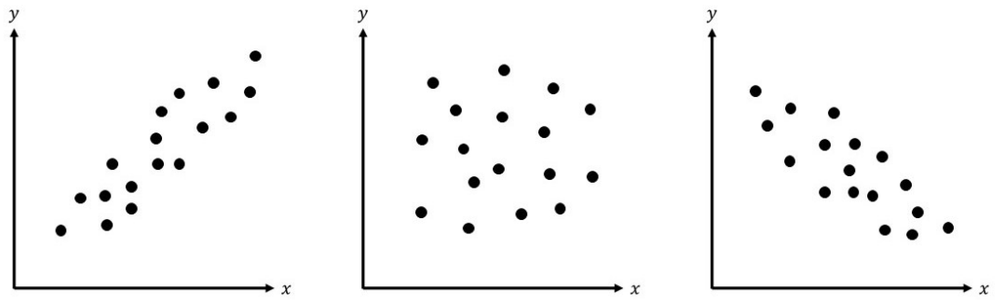
ちょうど, $\sin, \cos, \tan$ の正負の関係と同じような図になっていますが, あのイメージで大丈夫です.

さて, この図を踏まえたうえで, 共分散の式を見てみましょう.

$$(\text{共分散}) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y})$$

これは, **点 $(\overline{x}, \overline{y})$ の右上, もしくは左下にある点が多いほど大きく (よりプラスに) なり, 逆に左上, もしくは右下にある点が多いと小さく（よりマイナスに）なることがわかります.**

そして, 先ほどの図を見てみると,



正の相関　　相関なし　　負の相関

これと対応させると,

正の相関 共分散は大きい（プラス）
相関なし 共分散はほぼ 0
負の相関 共分散は小さい（マイナス）

となることがわかります. これが共分散なのですが, あれ？これで本当にいいの？と思った方がいたとしたら, とても鋭い視点を持っています. **実は, 相関を示す指標として共分散は不完全なのです.** 詳しくは次節で解説します. [*10]

相関係数

先程少し匂わせた共分散の問題点, 感づいたかたはいらっしゃるでしょうか？実は, それを解決するのが相関係数なのです.

と言われても, まだ共分散の弱点がピンときてない人もいるかと思いますので, まずはそちらをチェックします.

共分散の弱点

先程, 何気なく

正の相関 共分散は大きい（プラス）
相関なし 共分散はほぼ 0
負の相関 共分散は小さい（マイナス）

と書きましたが, ここでいう大きい, 小さいとはどれくらいを指すのでしょうか？共分散が 10 とかが相場なら,

[*10] すぐに解説してしまうので, 自分で考えてから進むのもありです.

100 というのは大きい（正の相関）と考えられますが, 1000000 とかが相場なら, 100 は 0 に近い (相関無し) と判定するべきです.

そして, この大きさの相場というものは, **定められません.** なぜなら, 例えば共分散が s_0 となるデータがあったとして, そのデータの x, y をともに 10 倍してあげる ($\{(1,2),(3,4)\cdots\}$ だったとしたら $\{(10,20),(30,40)\cdots\}$) と,

$$s_0 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y})$$

に対して, 新しいデータの共分散 s は

$$\begin{aligned} s &= \frac{1}{n}\sum_{k=1}^{n}(10x_k - 10\overline{x})(10y_k - 10\overline{y}) \\ &= \frac{1}{n}\sum_{k=1}^{n}10(x_k - \overline{x})\cdot 10(y_k - \overline{y}) \\ &= 100\frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})\cdot(y_k - \overline{y}) \\ &= 100\cdot s_o \end{aligned}$$

と, s_0 **の 100 倍になってしまいます！**しかし 10 倍しただけではデータの相関は変わっていないはずなので, これはゆゆしき事態だということがわかります.

ということで, 大きさの判定が難しいという弱点を共分散は持っているのです！

相関係数の導入

以上の弱点を克服したのが相関係数です. ただ, 定義がちょっといかついです.

相関係数の定義

$$\begin{aligned} (\text{相関係数 } r) &= \frac{(\text{共分散 } s_{xy})}{(x \text{ の標準偏差 } s_x)(y \text{ の標準偏差 } s_y)} \\ &= \frac{\sum(x_k - \overline{x})(y_k - \overline{y})}{\sqrt{\sum(x_k - \overline{x})^2}\sqrt{\sum(y_k - \overline{y})^2}} \end{aligned}$$

分母は (x_k, y_k) という組を無視し, x_k のみ, 及び y_k のみによって算出されます. 一方この時の分子は, $\frac{1}{n}$ が無いとて基本は共分散ですので, 正の相関が強いと大きく, 負の相関が強いと小さくなります.

ただ, ここで大切なのは, **相関係数** r **は** $-1 \leq r \leq 1$ **を満たすということです.** (← **超重要！**) すなわち,

正の相関 相関係数は 1 に近い
相関なし 相関係数は 0 に近い
負の相関 相関係数は-1 に近い

と, とても明確な基準にできるのです. 素晴らしい.

あんまりピンと来ない人も多いかと思うので, コーナーケースとして, もっとも相関が強い, すなわち $x_k = y_k$

ならば

$$\begin{aligned} r &= \frac{\sum(x_k-\overline{x})(y_k-\overline{y})}{\sqrt{\sum(x_k-\overline{x})^2}\sqrt{\sum(y_k-\overline{y})^2}} \\ &= \frac{\sum(x_k-\overline{x})(x_k-\overline{x})}{\sqrt{\sum(x_k-\overline{x})^2}\sqrt{\sum(x_k-\overline{x})^2}} \\ &= \frac{\sum(x_k-\overline{x})^2}{\sum(x_k-\overline{x})^2} \\ &= 1 \end{aligned}$$

となり, 逆に一番負の相関が強い $y_k=-x_k$ の場合では

$$\begin{aligned} r &= \frac{\sum(x_k-\overline{x})(y_k-\overline{y})}{\sqrt{\sum(x_k-\overline{x})^2}\sqrt{\sum(y_k-\overline{y})^2}} \\ &= \frac{\sum(x_k-\overline{x})(-x_k-(-\overline{x}))}{\sqrt{\sum(x_k-\overline{x})^2}\sqrt{\sum(-x_k-(-\overline{x}))^2}} \\ &= \frac{-\sum(x_k-\overline{x})(x_k-\overline{x})}{\sqrt{\sum(x_k-\overline{x})^2}\sqrt{\sum(x_k-\overline{x})^2}} \\ &= \frac{-\sum(x_k-\overline{x})^2}{\sum(x_k-\overline{x})^2} \\ &= -1 \end{aligned}$$

となります.

ということで, 相関係数を用いれば, 相関の強弱についてうまく判断できることがわかりました. これで, 相関係数も理解できたかと思います.

ただ, なぜ r が $-1\leq r\leq 1$ がを満たすかは, ここでは述べていません. この証明はいい練習問題なので, 皆さんの課題としたいと思います. [*11]

6 本章で触れなかったところについて

6.1 分散の公式

分散は, 定義通りにも求められますが,

$$s_x^2=\overline{x^2}-(\overline{x})^2$$

という公式を用いることもできます.

ただ, この公式は直観的というよりは, 式変形により導かれるものなので, どちらかといえば暗記になってしまうため今回は付録としました. 受験では結構使えるので, 覚えておくといいと思います.

6.2 データの変換

変量 x に対して, 新しい変量 y を $y=ax+b$ $(a\neq 0)$ で定義すると, 平均は $\bar{x}\to\bar{x}+b$, 標準偏差は $s\to as$ になります.

昔は発展問題で, 出題されたときその場で考えればいいって感じだったのですが, 最近何故かセンター及び共通テストでよく出題されます. [*12]ということで, 知っておいて損はありません.

[*11] あんまり教科書とかで見かけませんね. なぜでしょうかね.

[*12] データの線形変換は割と実用上意味を持つからかもしれません. データサイエンスとかホットですね

6.3 分散の別の表現

分散のパートで, $\sum_{k=1}^{n}(x_k-\overline{x})$ が 0 になってしまうから, $\sum_{k=1}^{n}(x_k-\overline{x})^2$ としたとありましたが, このときに, $\sum_{k=1}^{n}|x_k-\overline{x}|$ でもいいんじゃね？と思った人がいたかもしれません. 実は, これは一種の流派で受け入れられている考え方で, ある種分散の別の表現ともいえるということを伝えておきます. [*13]

6.4 偏差値について

これは完全に受験ってことではないのですが, 皆さん大好き偏差値についての話題です. いわば教養です.

みんな偏差値を気にするくせに, 偏差値の定義を知らないのではないでしょうか. そんな人の偏差値は知れています. 是非ここで定義を覚えていってください.

偏差値の定義

$$(\text{偏差値}) = 50 + \frac{(\text{自分の点数}) - (\text{平均})}{(\text{標準偏差})} \times 10$$

偏差値はもともと, 平均が安定しないテストなどにおいても, 通常の点数のように善し悪しが判断できるよう考案されたもので, 50 というのは意味があるようでないのです.

注意すべき点としては,

- 平均と全く同じなら偏差値 50.
- 平均が満点に近いと高偏差値は出せない.
- 標準偏差が大きいと偏差値が高くなりにくい. （みんな同じくらいの点数の中高得点を取ると高偏差値が出る）

ということです. これでこれからはより高い偏差値を狙えますね！[*14]

7 終わりに

今回はデータの分析をまとめてみましたが, いかがだったでしょうか. 苦手, 嫌いな人が多いこの単元の学習の一助になれば幸いです.

それでは良いデータの分析ライフ（？）を.

[*13] この場合単位も狂わなくて便利です.

[*14] 期待しています.