https://www.math.sci.ehime-u.ac.jp/~matsuura/

データサイエンス入門

～ プロ野球で学ぶデータ解析 ～

愛媛大学　　松浦 真也

Ver. 2024.4.9

本資料について

本資料の著作権は放棄していませんが、教育機関にて、教育目的で本資料をご活用いただくことは、歓迎致します。ただし、データや画像等を引用している箇所は、引用元のルールを遵守願います。

- インターネット上などでの無断転載はご遠慮下さい。
- 生徒・学生の皆さんが宿題やレポート等に本資料を活用する際は、引用元の不記載など、引用の要件を満たさないコピペは厳に慎んで下さい。
- 記載内容の厳密性については、保証しかねます。非専門家向けに、平易で簡潔な説明を行うことを目的としていますので、厳密な説明は意図していません。特に、各種分析においては、厳密性よりデータ解析の楽しさが伝わることを優先しています。
- 筆者がプロ野球中日ドラゴンズのファンのため、ドラゴンズの選手やドアラ（マスコットキャラクター）のデータを中心に例示しています。
- 画像の一部は生成AI（Adobe Firefly）で作成しました（当該画像にはその旨記載）。
- 本資料の利用に伴う損害や不利益については、一切の責任を負いかねます。
- 本資料の一部または全部を、予告なく削除したり、修正したりする可能性があります。

本資料の内容

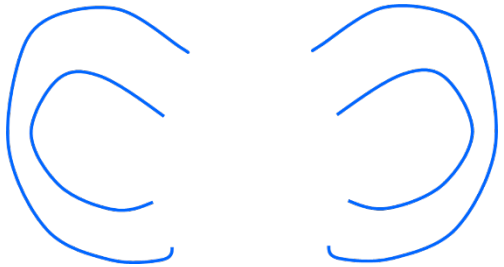
この資料で学ぶこと

- いろいろなデータ：数値データの他にも、テキストデータや画像データなどがある
- 疑似相関：2種類のデータの間に、相関関係（片方の値が大きいほど、もう片方の値も大きいという関係）があることと、因果関係（片方の値が、もう片方の値に影響を及ぼすという関係）があることとは、別物
- 線形回帰分析：2種類のデータの散布図（片方の値を横軸に取り、もう片方の値を縦軸に取ったグラフ）を、直線で近似（3種類以上のデータにも拡張可能）
- 非線形回帰分析：2種類のデータの散布図を曲線で近似
- 主成分分析：多くの項目からなるデータの特徴を要約し、平面上に可視化
- 統計的なバラツキ：確率現象においては、条件が一定でも、結果にバラツキが生じる。そうした、単なる「偶然の産物」に対して、特別な「意味」を見いださないように留意が必要

ドアラのデータを分析してみよう！

いろいろなデータ

～ ビジネスへの活用も念頭に ～



ドアラの耳

いろいろなデータ

そもそも、データって何？

一言で正確に説明するのは難しいが、大雑把には、

分析や解釈の対象となり得る基礎的な客観的事実。

※ データは、数値、テキスト（文章）、画像、音声などの形で表現される。

「データ」と「情報」の違い

データは、それ自体は直ちに有益な意味を持つとは限らない。

情報は、データを整理・分析・解釈することで、有益な意味を持たせたもの。

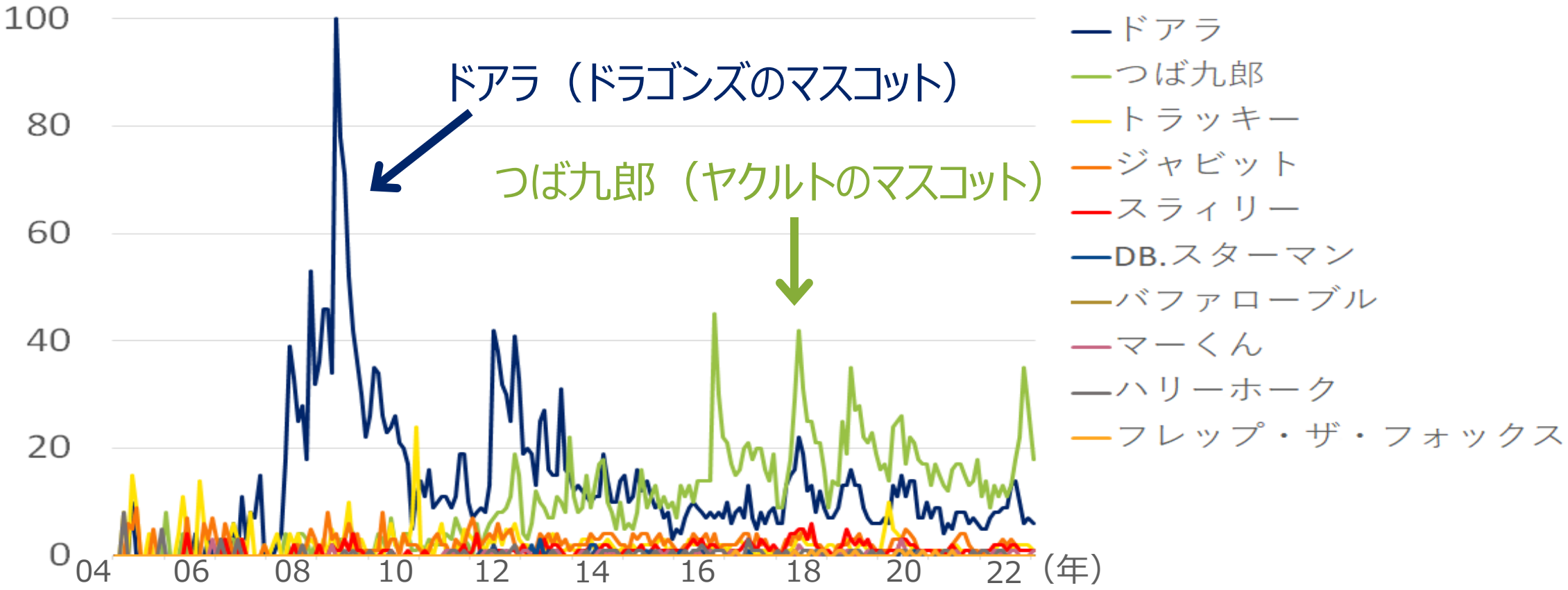
堅い話はこれでおしまい。

以後、数値データ、テキストデータ、画像データの具体例を見てみよう！

数値データ

数値データの例　球団マスコットの検索回数の推移（2004.4～2022.1）

※期間中の最大検索回数を100とし、検索回数の多さ（相対的な値）をグラフ化



データ出典：Google Trends https://trends.google.co.jp/trends/ 2022年1月22日検索

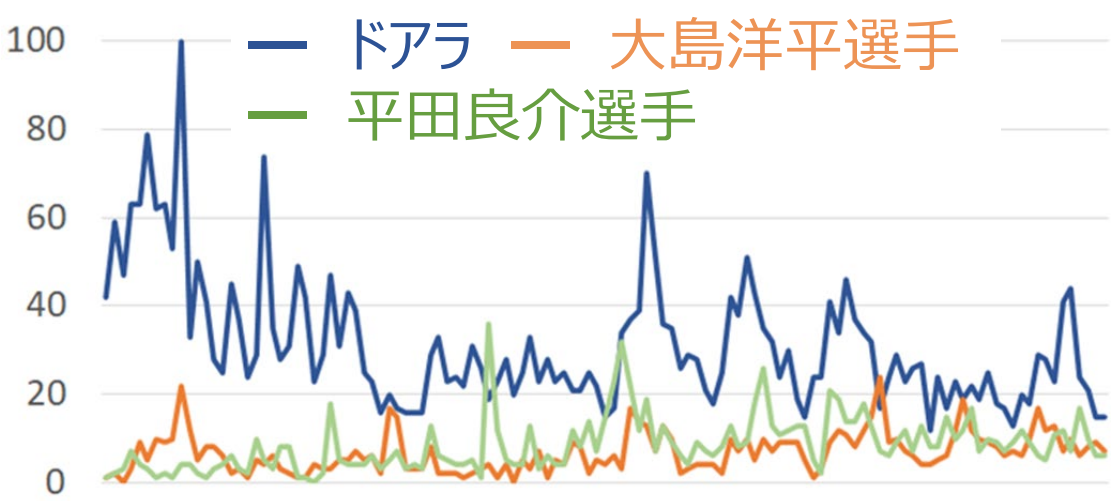
数値データ

数値データの例

ちょっとだけ、ビジネスの話

中日ドラゴンス

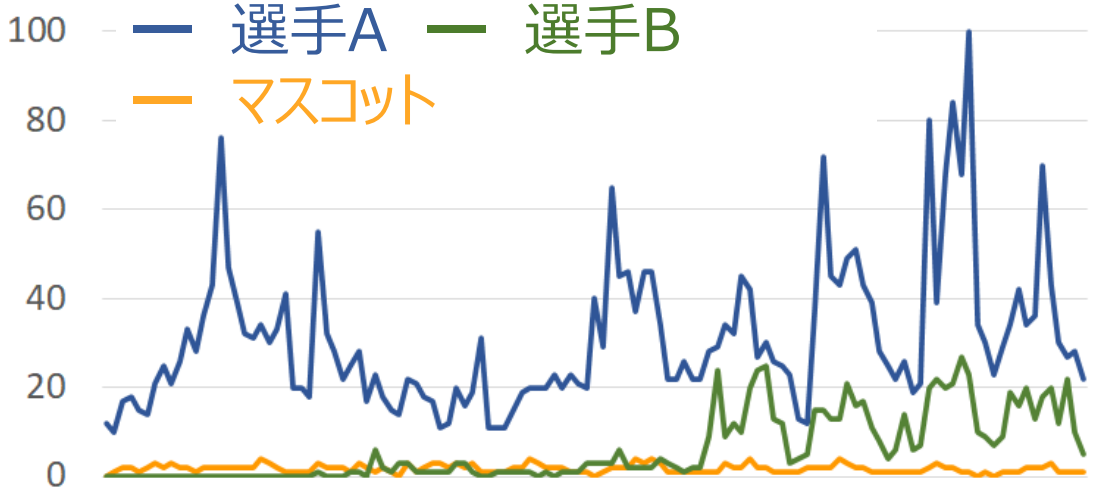
ドアラが根強い人気



ドアラ関連の商品を拡充！

他球団の例

選手が根強い人気



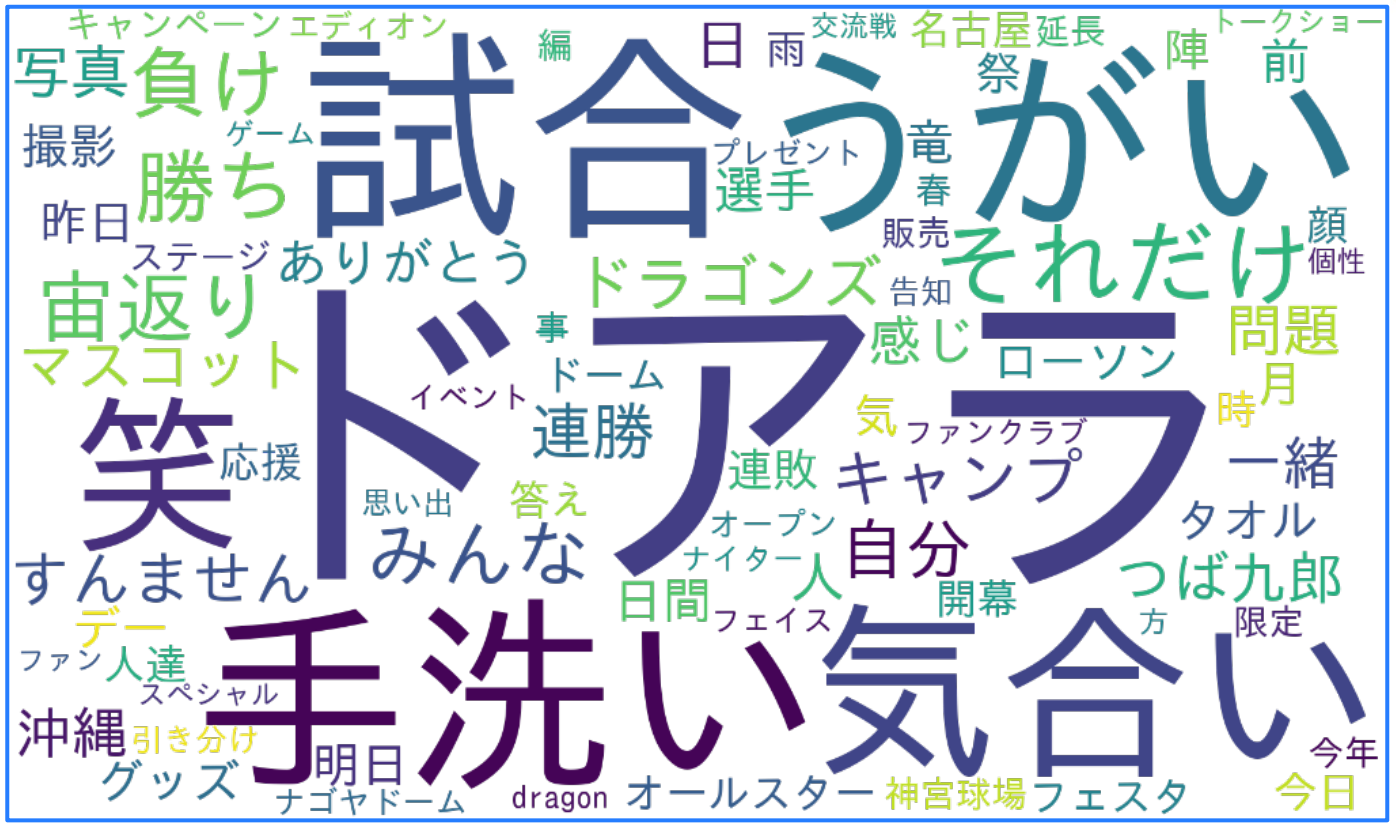
選手A・B関連の商品を拡充！

データ出典：Google Trends https://trends.google.co.jp/trends/ 2022年1月22日検索

テキストデータ

テキスト（文章）データの例　中日ドラゴンズ ドアラ 公式ブログ

ドアラのブログのワードクラウド



ドアラが書き込んだ約16万字（2016.12～2022.1の分）をプログラミング言語Pythonで解析

↓

ワードクラウドを作成

頻度・重要度の高い単語を強調して図示

- ドアラは、つば九郎と仲良し。
- うがい、手洗いなど、時代を反映。

※データ・写真の利用について、株式会社中日ドラゴンズ様から許諾を得ています。転載はご遠慮下さい。

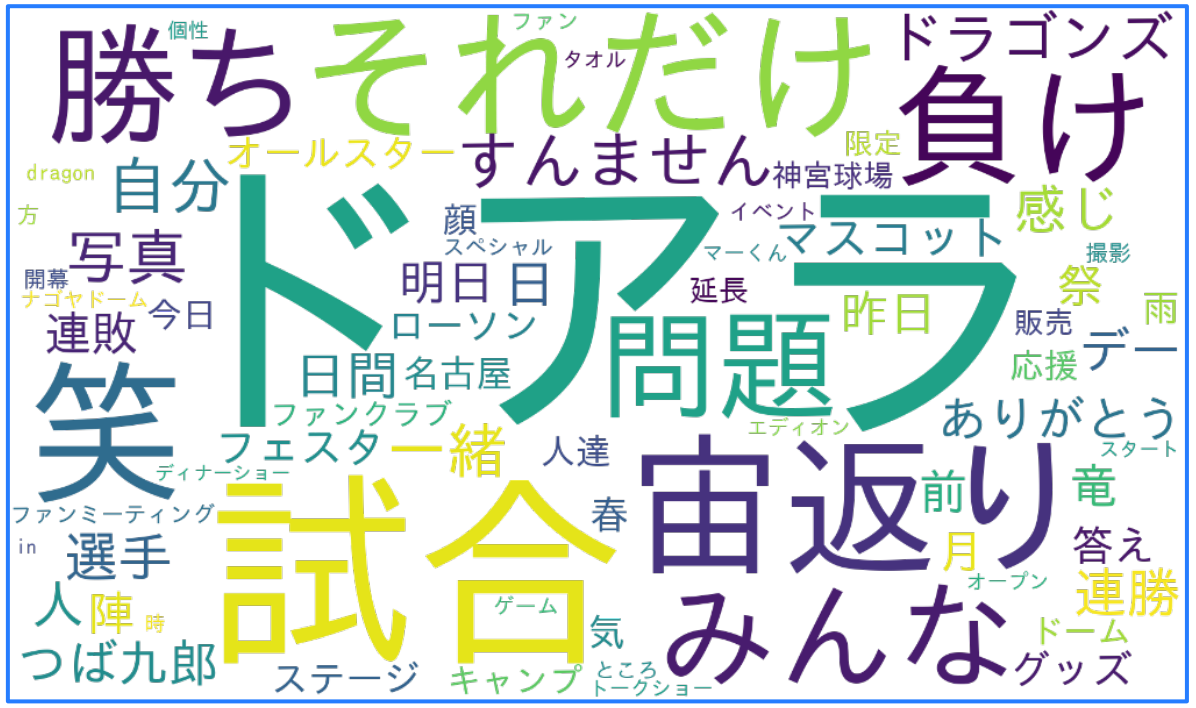
テキストデータ・写真出典：ドアラ公式ブログ　https://profile.ameba.jp/ameba/doalaofficial/

テキストデータ

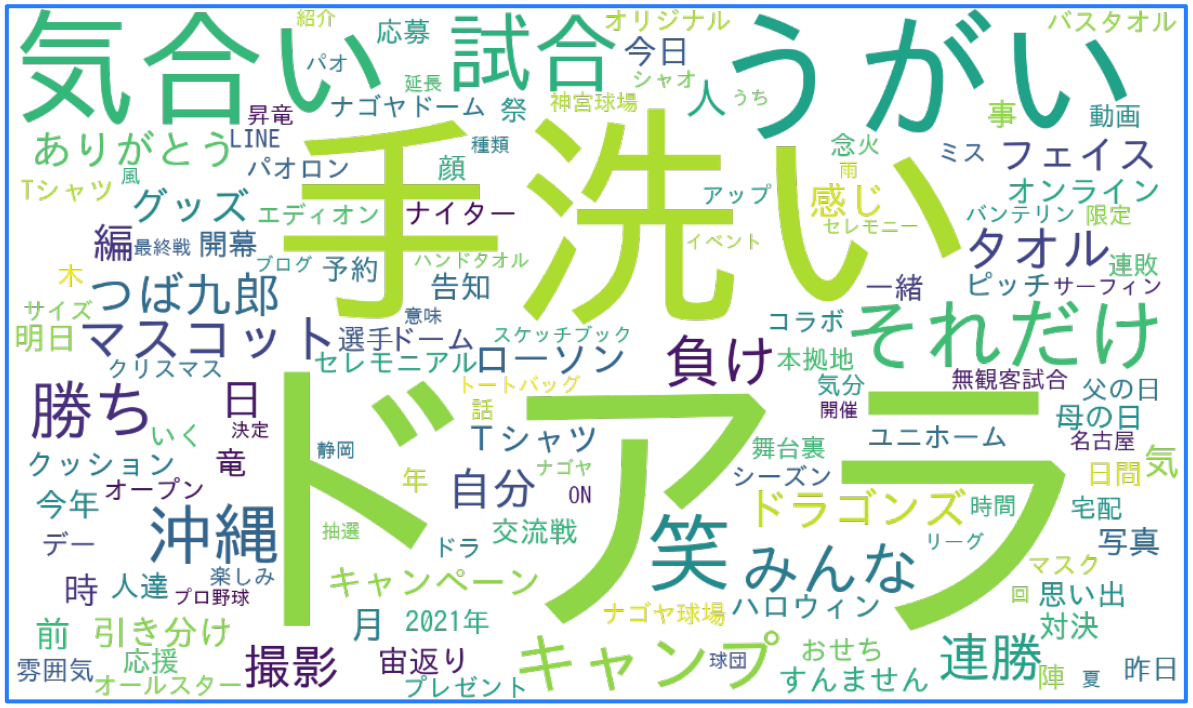
テキスト（文章）データの例

ドアラ公式ブログ COVID-19（新型コロナウイルス感染症）前後の比較

COVID-19前（2016.12～2020.1）



COVID-19後（2020.2～2022.1）



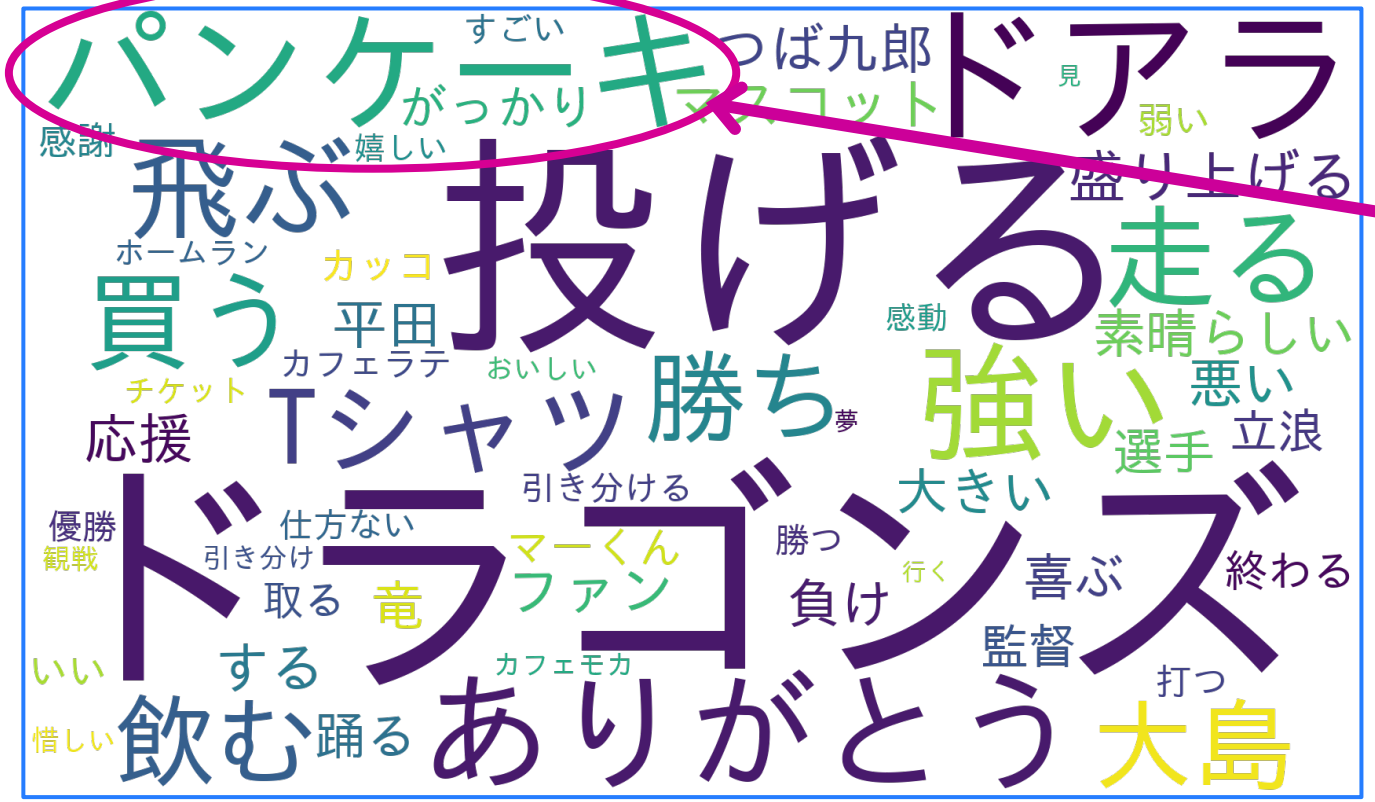
マスコット界隈も、COVID-19の影響を受けた模様

- COVID-19前に盛んだったファンとの交流や遠征が、COVID-19後は減少
- COVID-19後は、「手洗い、うがい、気合」を呼びかけ
- つば九郎（ヤクルトのマスコット）との友情は不変！

テキストデータ

テキスト（文章）データの例　ちょっとだけ、ビジネスの話

ドラゴンズファン達のSNSへの書き込み（架空のデータ）



テキストマイニング
テキストデータから宝探し

ドラゴンズと無関係？
でも、無関係そうなものほど、お宝！
（意外な組み合わせが、ヒットを生む）

ドアラの似顔絵が描かれた
ドアラ・パンケーキを開発

画像データ

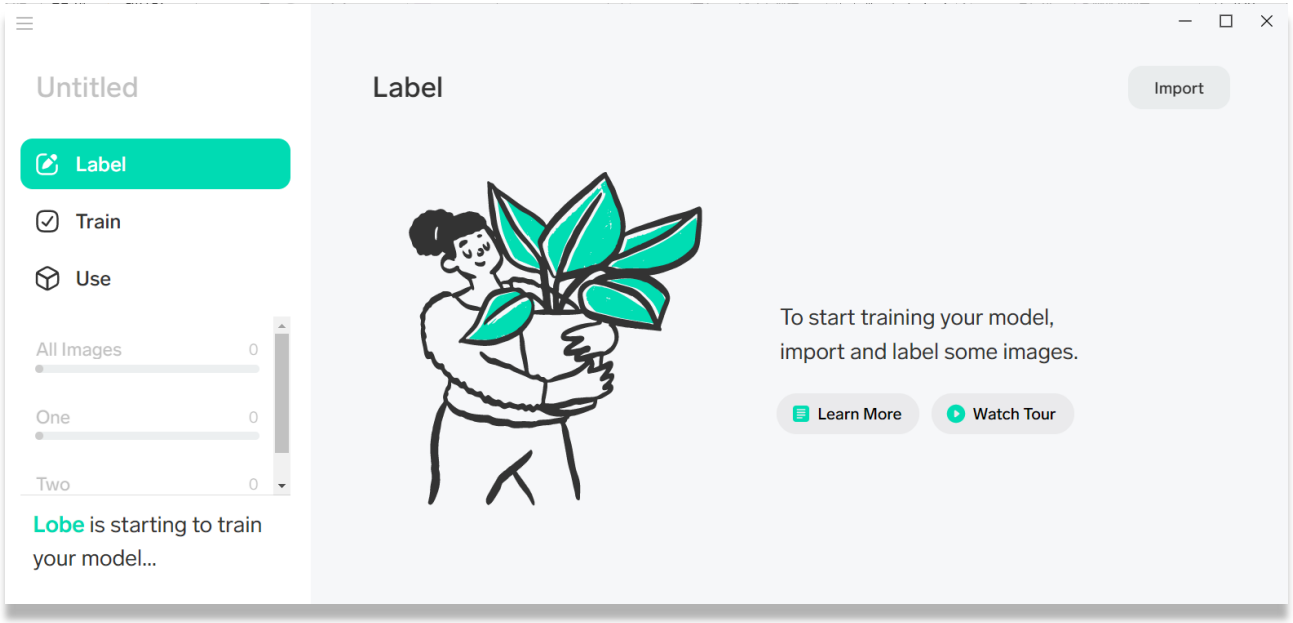
画像データの例　犬猫判定

機械に勉強させ、画像に写った動物が、犬か猫かを自動で判定。

専用サイト、アプリも存在し、訓練用画像を読み込ませ、ボタンを押すだけ！

※犬と猫でなくても、何の画像でも可

※Python等でも実装可



判定アプリの例：lobe　https://www.lobe.ai/

画像：生成AI（Adobe Firefly）により作成

画像データ

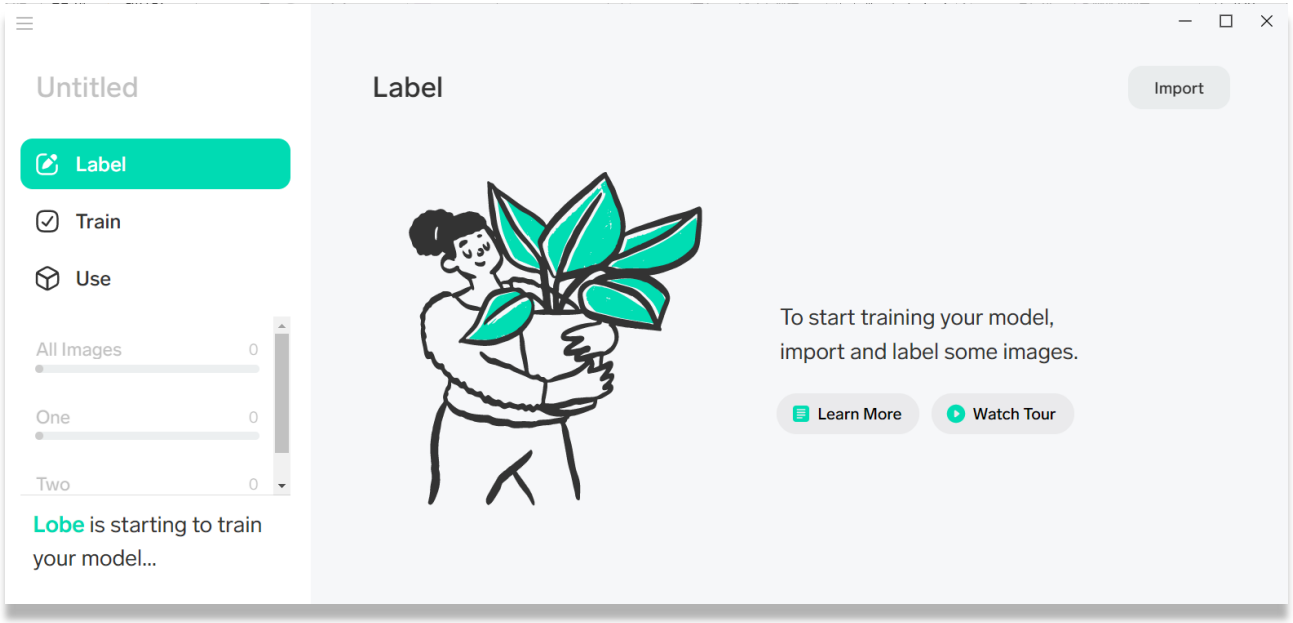
画像データの例　ドア九郎判定

機械に勉強させ、画像に写った動物が、犬か猫かを自動で判定。

専用サイト、アプリも存在し、訓練用画像を読み込ませ、ボタンを押すだけ！

※犬と猫でなくても、何の画像でも可　→　画像に写ったマスコットが、ドアラかつば九郎か自動で判定

※Python等でも実装可



判定アプリの例：lobe　https://www.lobe.ai/

画像：生成AI（Adobe Firefly）により作成

画像データ

画像データの例　ちょっとだけ、ビジネスの話

異常検出

パンケーキの焼け具合をカメラで常時監視

↓

「正常な画像」か「異常な画像」か判定
（ドア九郎判定と同じ原理）

↓

リアルタイムに異常を自動検出

ドアラのパンケーキ工場

画像：生成AI（Adobe Firefly）により作成

ヒットを打ちたければ、どんどん三振せよ？？

疑似相関

～ 見せかけの関係性に注意 ～

ヒット・三振数（使用データ）

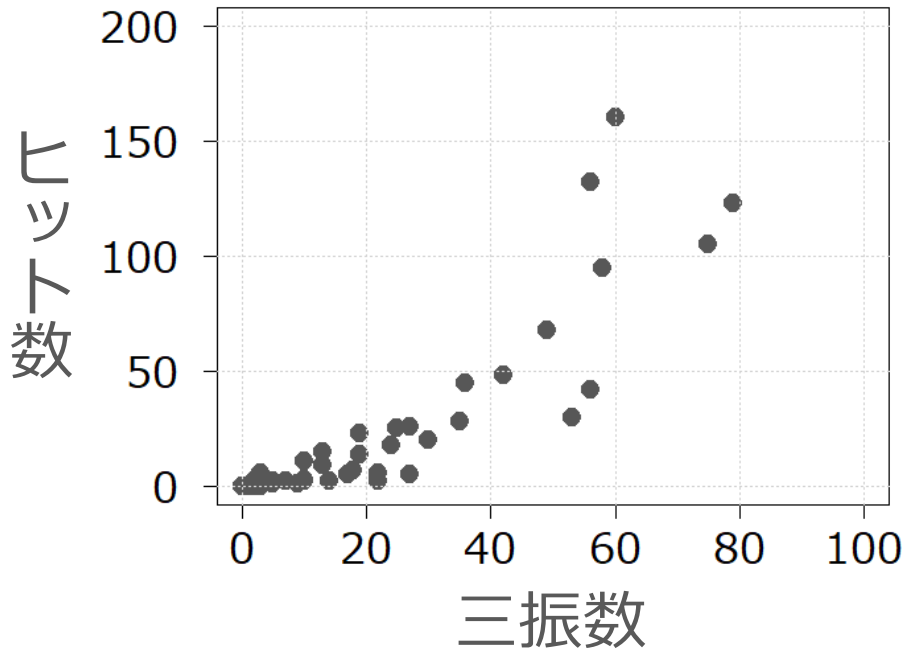
三振とヒット（安打）の数

ここでは、投手（ピッチャー）も含め、2021年に、中日ドラゴンズに所属した各選手の2021年シーズンにおける一軍公式戦でのヒット数と三振の数のデータを分析する。

分析対象のデータ

背番号	選手	三振数	ヒット数
0	高松	35	28
00	石岡	5	1
1	京田	75	105
3	高橋周	79	123
以下略			

全60選手分のデータをグラフ化



データ出典：中日ドラゴンズ公式HP シーズン打撃成績（2022年2月12日閲覧）
※現時点では、直近のシーズンのデータに更新済
https://dragons.jp/teamdata/batting.html

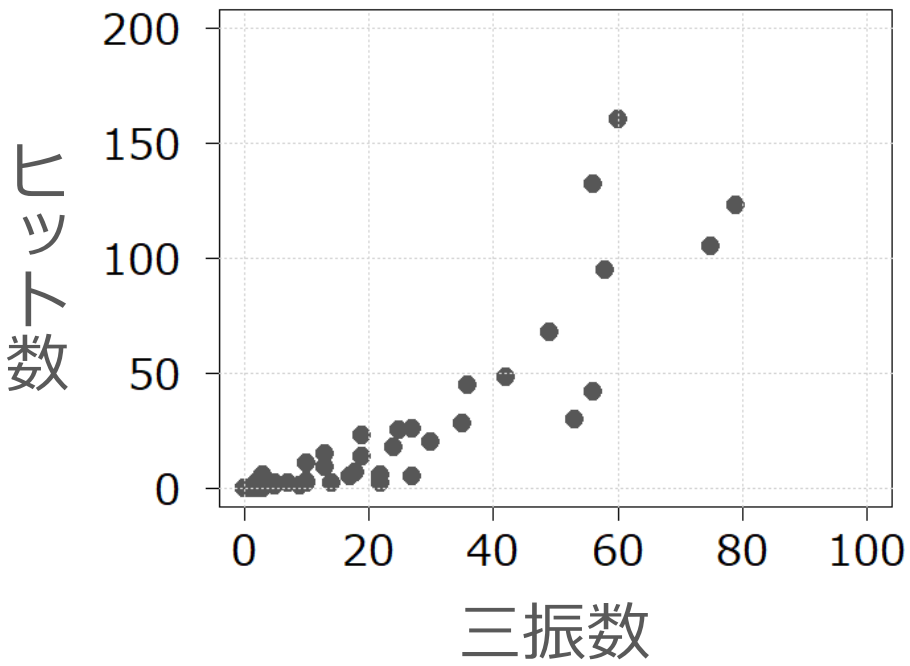
ヒット・三振数（相関関係）

ヒットを増やすには？

グラフを見ると、三振の数が増えれば、ヒットの数も増える！

三振は「大振り（バットを振り回す）」というイメージ。

コンパクトにバットを振った方が、ヒットを多く打てそうだが、実は違う？



ヒットを狙いすぎると、思いきりが悪くなり、結果として、ヒットが打てないのかも。

ヒットを打ちたければ、どんどん大振りして三振しよう！

でも、ちょっと待って！！！

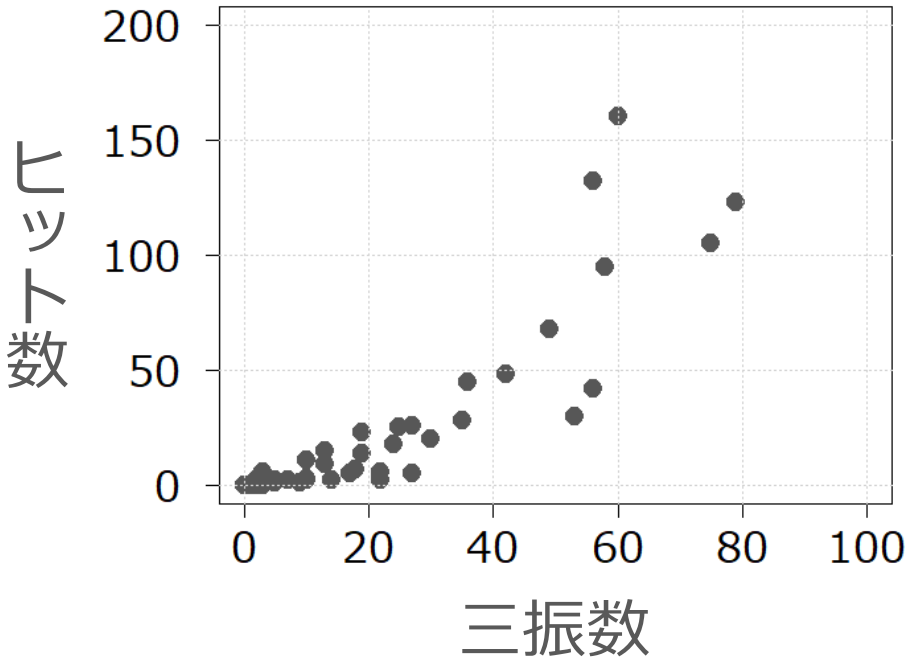
ヒット・三振数（相関関係）

ヒットを増やすと？

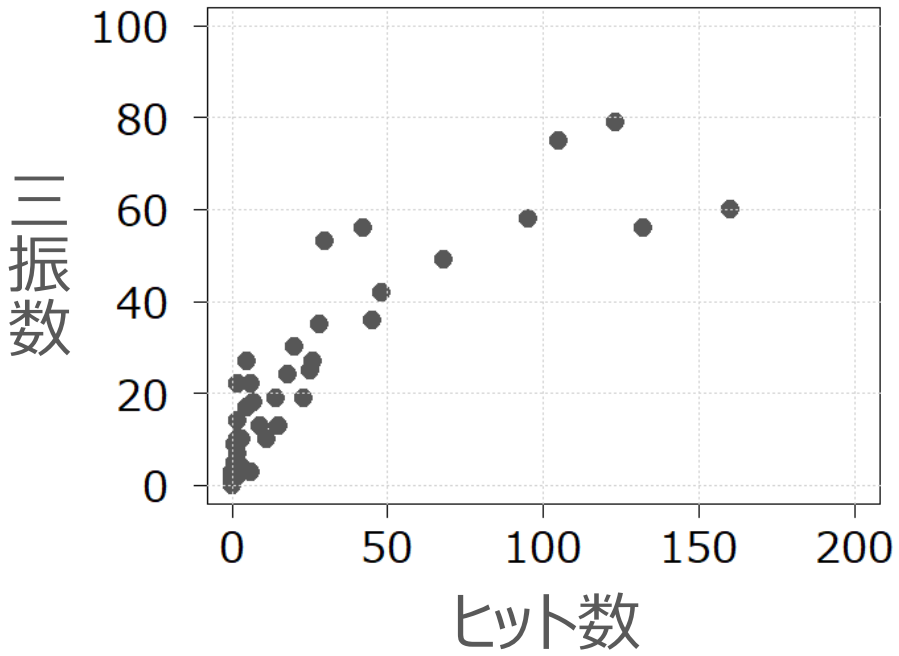
縦軸と横軸を入れ替えてみると …… ヒットの数が増えれば、三振も増える！

三振が先か、ヒットが先か？

三振を増やせば、ヒットが増える？



ヒットを増やせば、三振が増える？



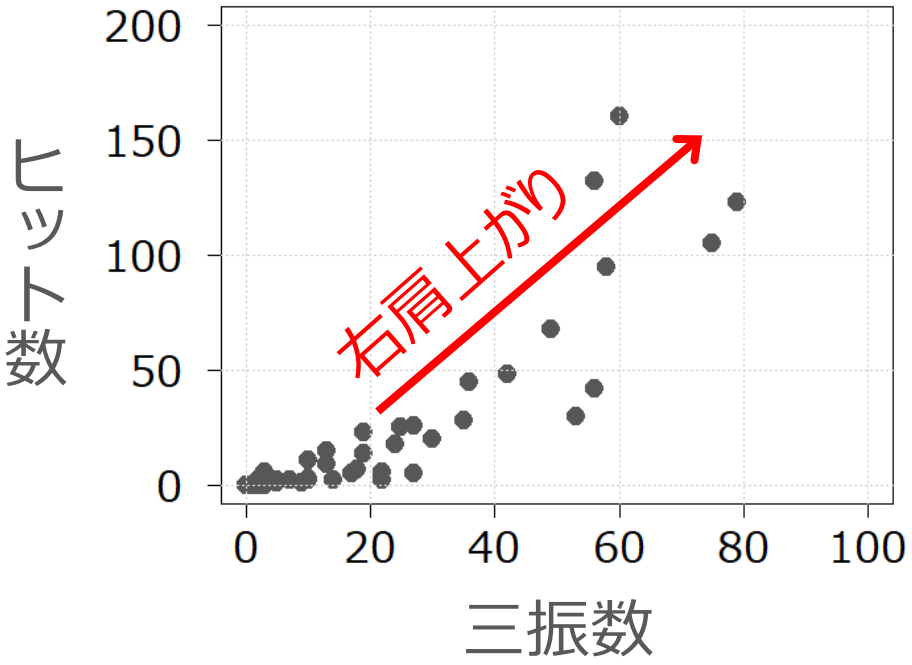
ヒット・三振数（相関関係）

相関関係

三振かヒットか、どちらが先かはともかく、
片方の値が大きいほど、もう片方の値も大きい傾向にある。
グラフで見ると、右肩上がりになっている。
このような関係にあるとき、両者には
「正の相関がある」という。

逆に、片方の値が大きいほど、もう片方の値は小さい
（グラフで見ると、右肩下がりの）場合は、
「負の相関がある」という。

正の相関がある



ヒット・三振数（疑似相関）

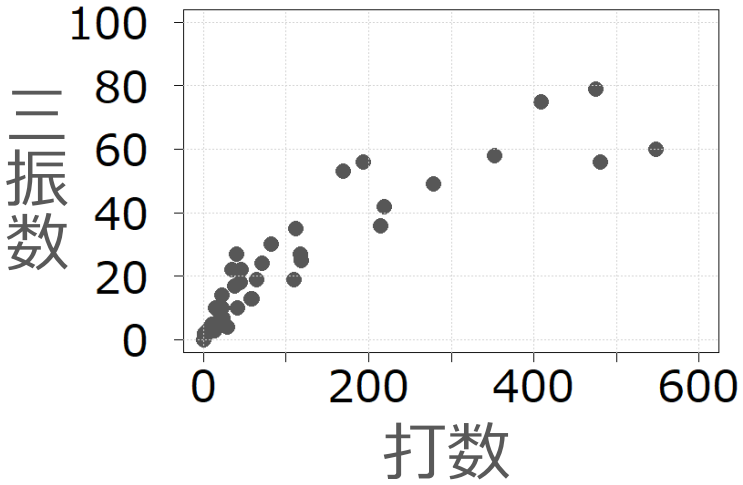
第3の要因　三振が先でもないし、ヒットが先でもない。根本的な要因は打数。

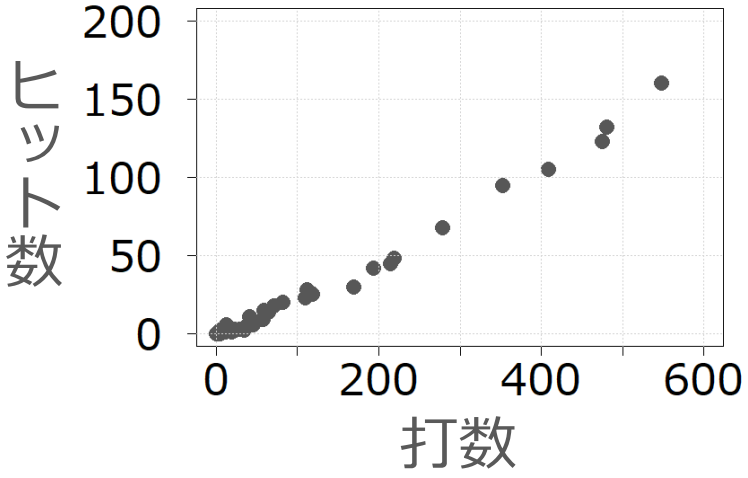
※打数とは、打席数（バッターボックスに立った数）から、四死球、犠打、犠飛、打撃妨害、走塁妨害の数を引いた数

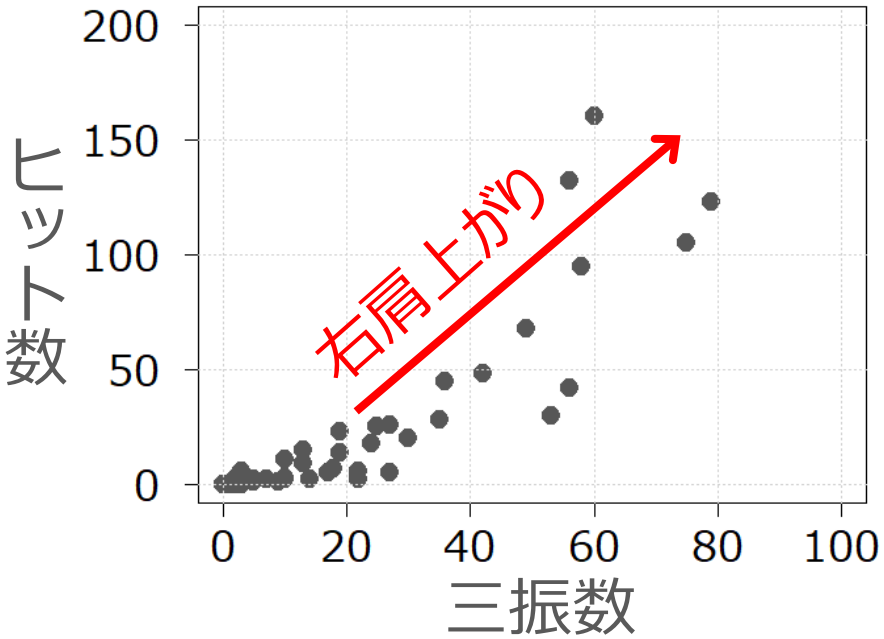
打数が増えれば、三振数が増える。打数が増えれば、ヒット数も増える。

結果として、三振の多さとヒットの多さは連動する。

このように、第3の要因（打数）を通じて、見せかけの関係性があるとき、両者に「疑似相関がある」という。





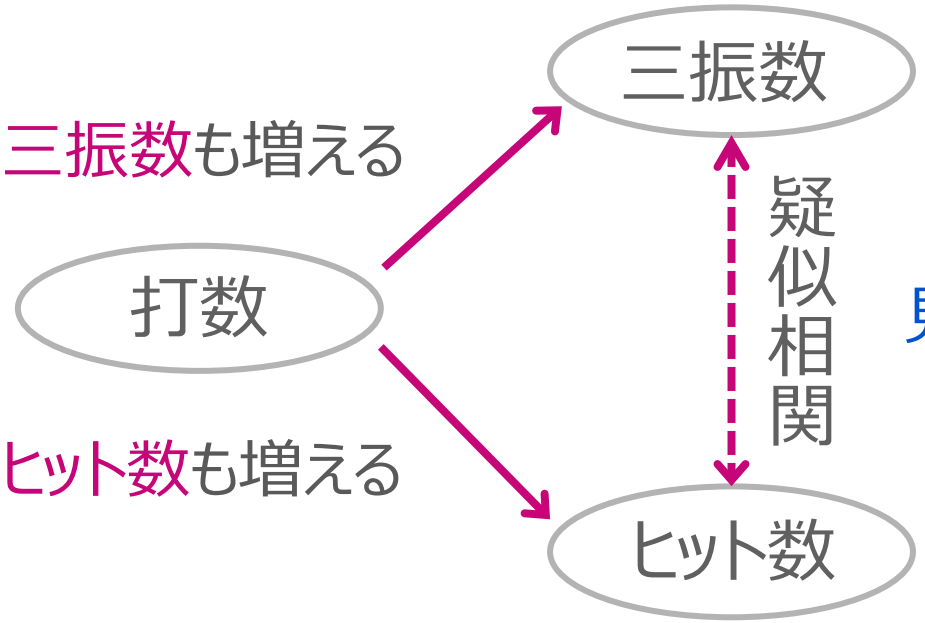


ヒット・三振数（疑似相関）

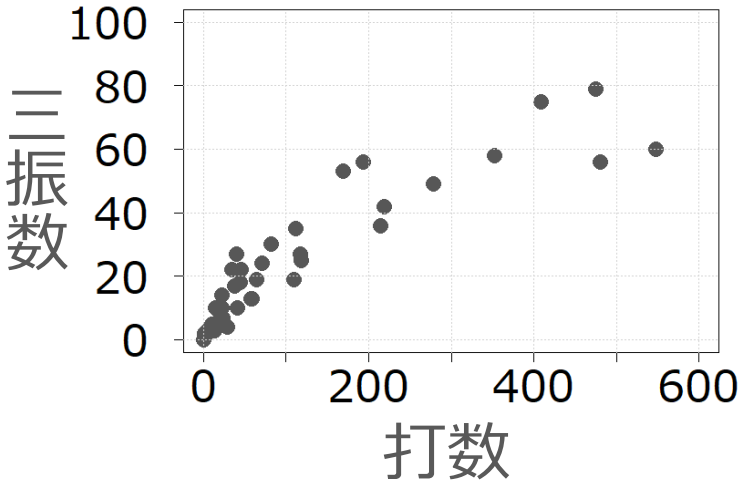
疑似相関

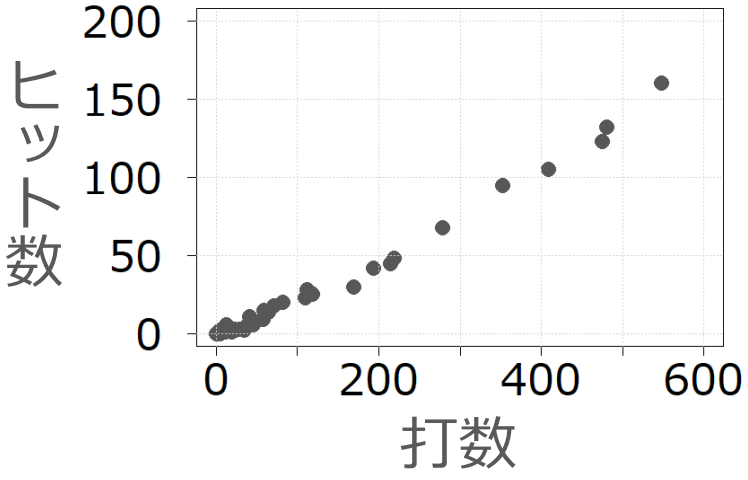
打数が増えれば、三振数も増える

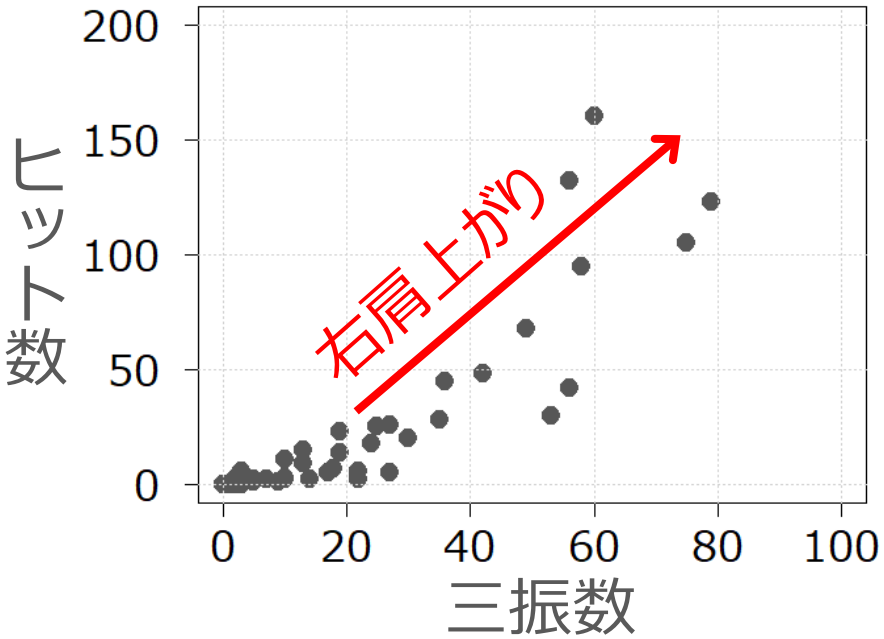
打数が増えれば、ヒット数も増える



見せかけの関係性が生じる







ヒット・三振数（相関関係再考）

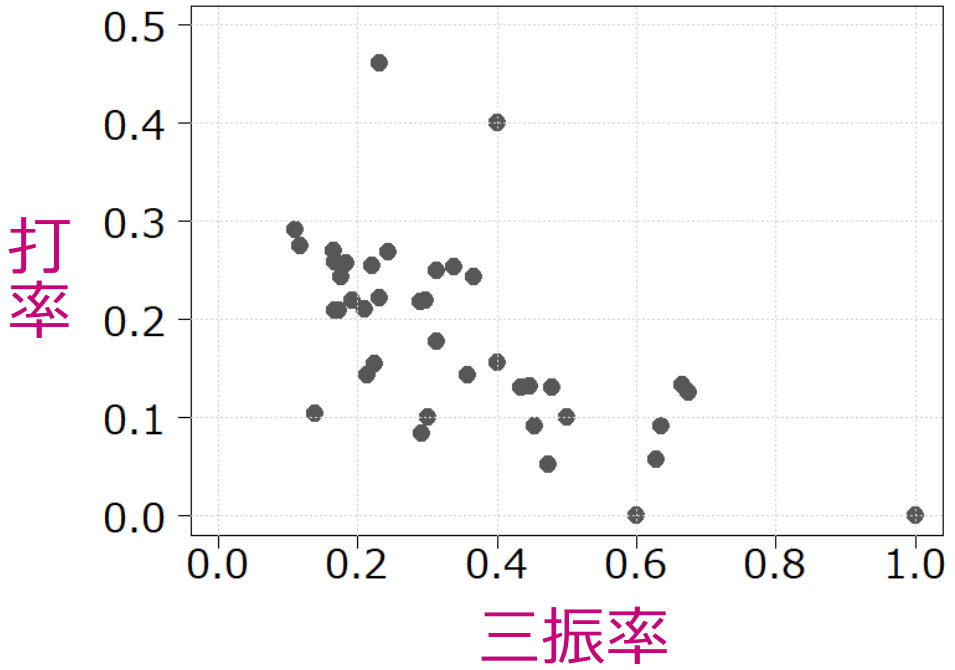
数ではなく率（割合）で考察

打数の影響を取り除くため、三振数とヒット数ではなく、
三振率（三振数÷打数）と打率（ヒット数÷打数）で考える。

分析対象のデータ

背番号	選手	三振率	打率
0	高松	0.313	0.250
00	石岡	0.455	0.091
1	京田	0.183	0.257
3	高橋周	0.166	0.259
以下略			

打数1以上の
全43選手分の
データをグラフ化



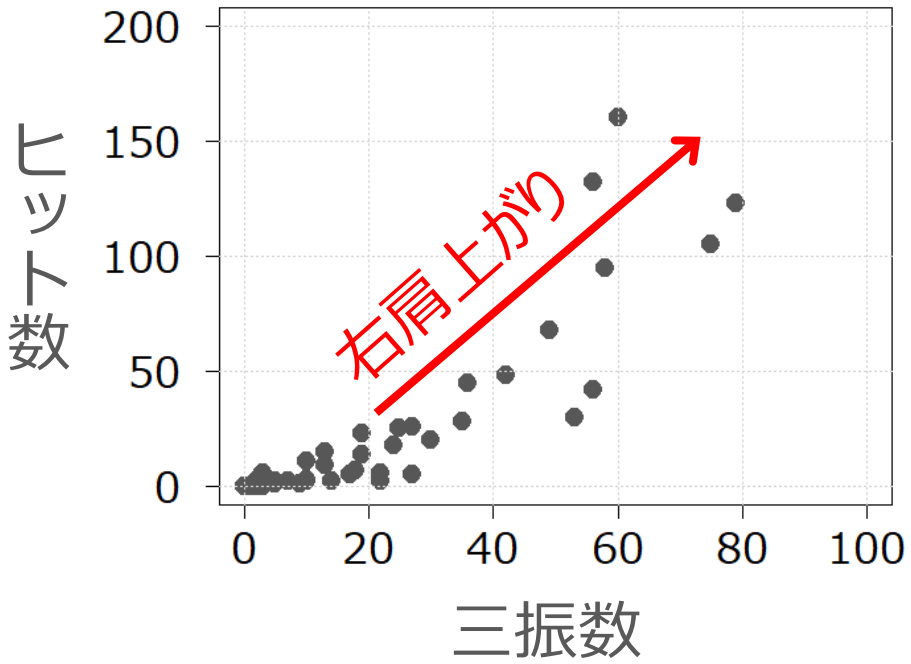
データ出典：中日ドラゴンズ公式HP シーズン打撃成績（2022年2月12日閲覧）から算出
※現時点では、直近のシーズンのデータに更新済
https://dragons.jp/teamdata/batting.html

ヒット・三振数（相関係数）

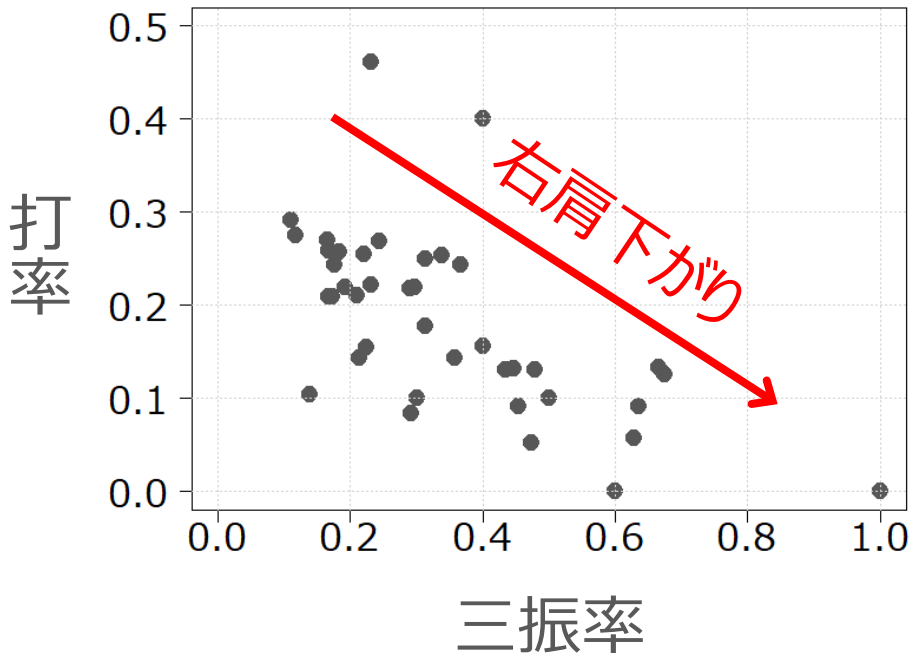
相関係数

相関係数とは、2種類のデータの関係性を表す数値で、－1以上1以下の値をとる。
強い正の相関がある（片方が増えると、もう片方も増える）とき、相関係数は1に近い。
強い負の相関がある（片方が増えると、もう片方は減る）とき、相関係数は－1に近い。

正の相関（相関係数0.880）



負の相関（相関係数－0.715）



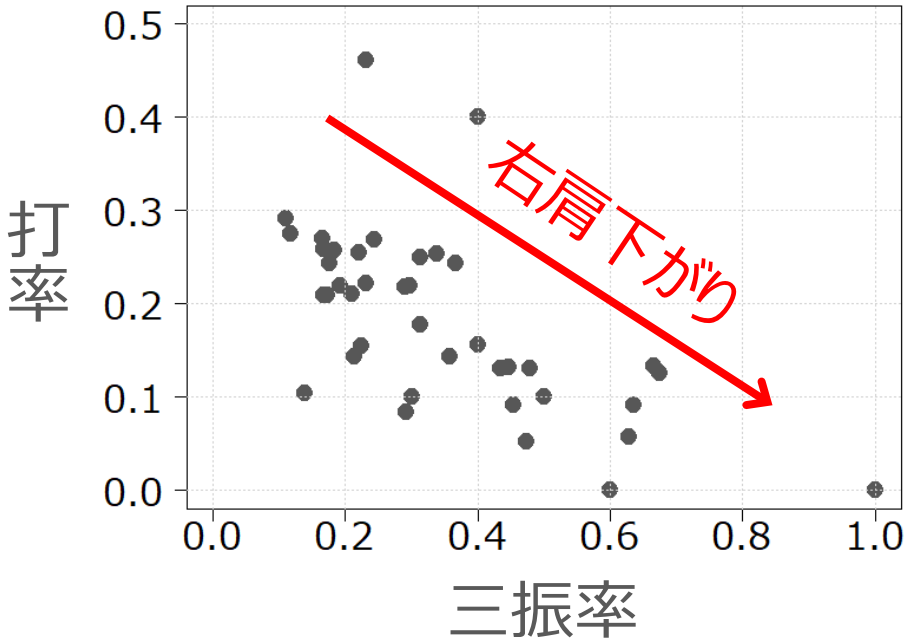
ヒット・三振数（まとめ）

三振とヒット（安打）の数（結論）

三振が多い選手は、ヒットも多い傾向にある（正の相関）。
しかし、「三振を増やせば、ヒットも増やせる」という因果関係があるわけではない。
三振が多い選手は、それだけ打数が多く、ヒットを打つチャンスも多い。

つまり、打数を通じ、三振数とヒット数の間に、
見せかけの関係が生じている（疑似相関）。
打数の影響を取り除くため、三振数とヒット数を
打数で割ったもの（三振率と打率）を考える。

相関係数　−0.715



結論：三振率と打率には負の相関がある

三振率が低い（三振しにくい）選手は、
打率が高い（ヒットを打ちやすい）傾向にある。

バッターの成績とチーム得点の関係は？

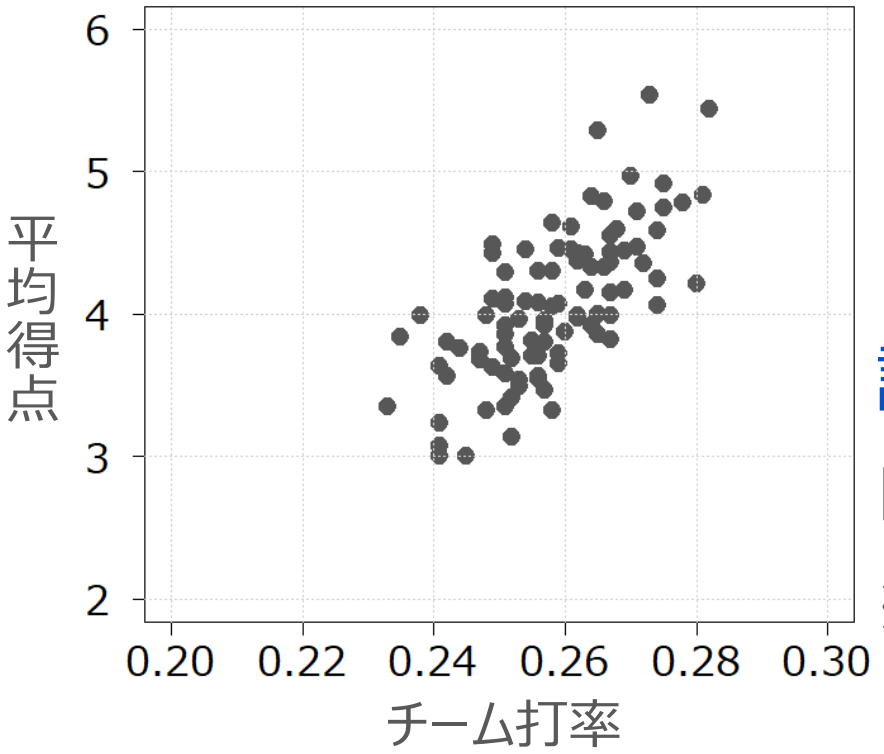
線形回帰分析

～ データを直線で近似 ～

平均得点（散布図）

2005～2020年プロ野球（パ･リーグ）チーム成績の回帰分析

散布図



標本サイズ（点の数）：96（6チーム × 16シーズン）
横軸：各年･各チームのチーム打率（ヒットを打つ確率）
縦軸：各年･各チームの1試合当たりの平均得点

試合の得点は、チーム打率と、どのような関係にある？

「チーム打率」を横軸、「平均得点」を縦軸に取り、散布図を描く。

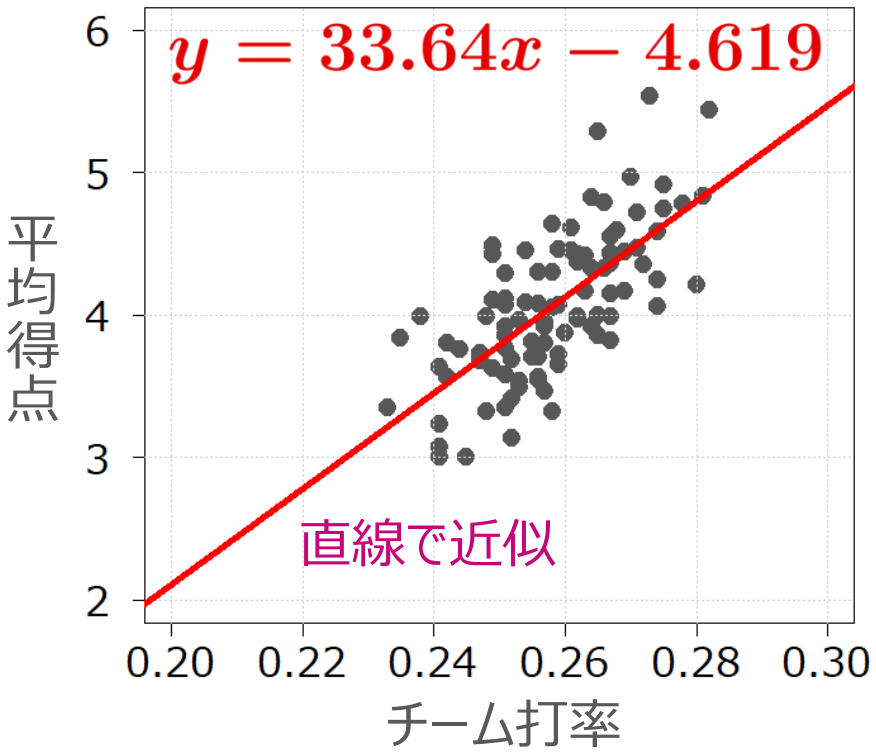
※ 一般に、ある2項目（この例では、打率と得点）からなるデータについて、片方の項目を横軸、もう片方の項目を縦軸に取り、データに対応する箇所に点を記した図を散布図という。

データの出典：パ･リーグ.com https://pacificleague.com/（2021年9月27日閲覧）

平均得点（線形単回帰分析）

2005～2020年プロ野球（パ･リーグ）チーム成績の回帰分析

単回帰分析



※横軸をx、縦軸をyとする。

標本サイズ（点の数）：96（6チーム × 16シーズン）
横軸：各年･各チームのチーム打率（ヒットを打つ確率）
縦軸：各年･各チームの1試合当たりの平均得点

大雑把には、右肩上がりになっているので、打率が高いと、得点力も高い（正の相関がある）。

点の分布を直線で近似することで、打率がどの程度だと、得点力がどの程度かを推計するための計算式を導出。

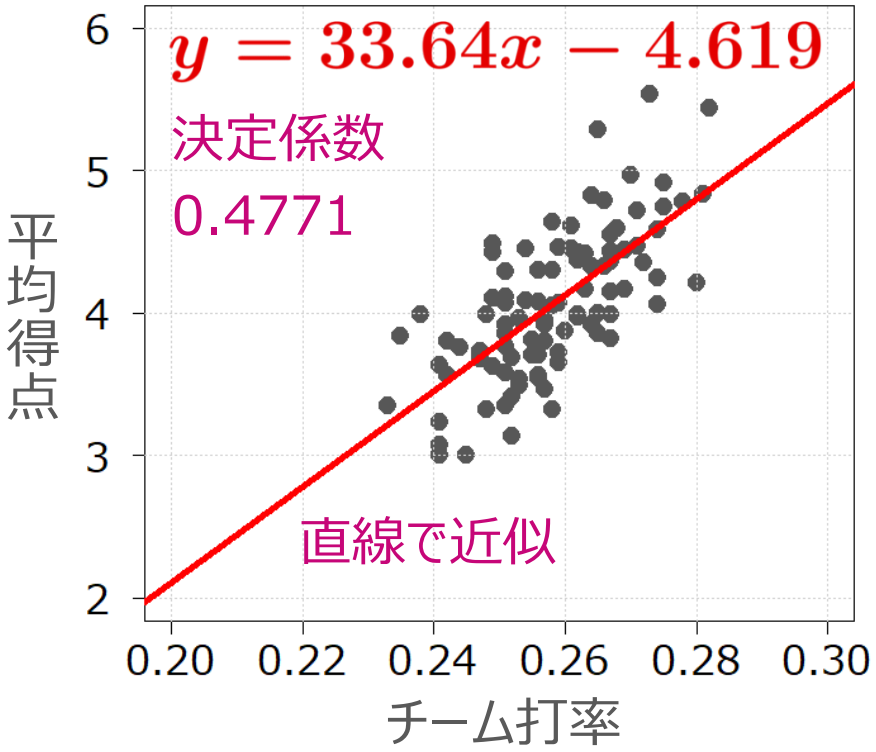
このように、点の分布を直線で近似することを、線形の単回帰分析という。　※「線形」は「直線的」の意味

データの出典：パ･リーグ.com https://pacificleague.com/（2021年9月27日閲覧）

平均得点（線形単回帰分析）

2005～2020年プロ野球（パ･リーグ）チーム成績の回帰分析

単回帰分析



※横軸をx、縦軸をyとする。

標本サイズ（点の数）：96（6チーム × 16シーズン）
横軸：各年･各チームのチーム打率（ヒットを打つ確率）
縦軸：各年･各チームの1試合当たりの平均得点

決定係数（大雑把な説明）：

点の分布を直線で近似したとき、近似しきれない部分（誤差）が残る。直線でどの程度近似可能か、その割合を示す値が決定係数。割合なので、0から1の値を取る。

決定係数が1に近い：ほぼ完全に直線で表せる
決定係数が0に近い：直線では、ほとんど何も近似できない

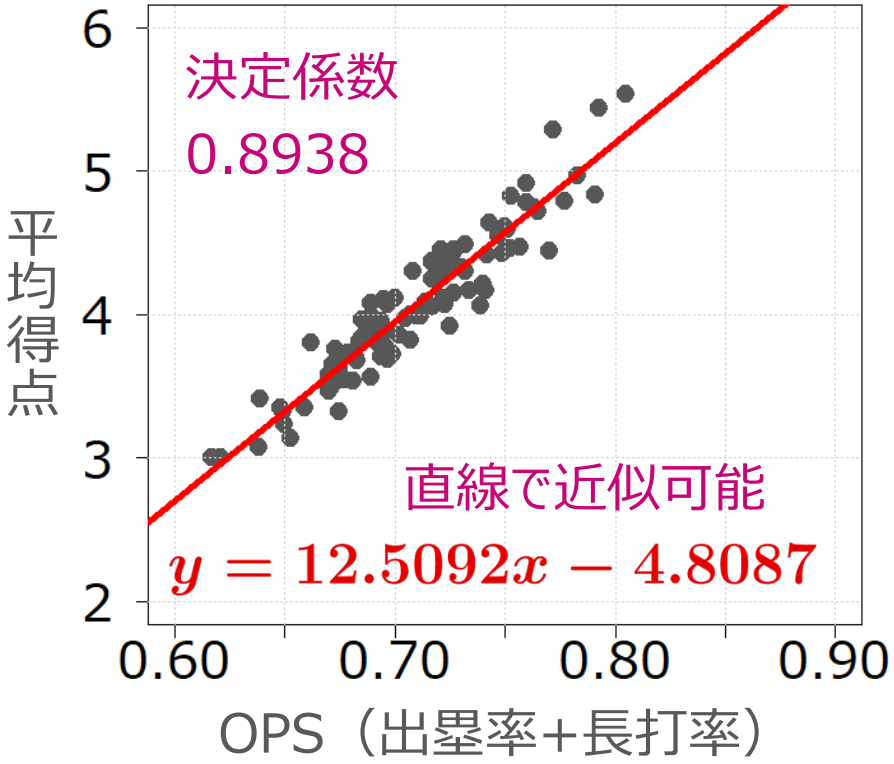
データの出典：パ･リーグ.com https://pacificleague.com/（2021年9月27日閲覧）

平均得点（線形単回帰分析）

2005～2020年プロ野球（パ･リーグ）チーム成績の回帰分析

実は、得点は打率よりOPS（出塁率＋長打率）を用いた方が、近似精度が良くなると言われている。

単回帰分析



$$\text{出塁率} = \frac{\text{ヒット数}+\text{四球数}+\text{死球数}}{\text{打数}+\text{四球数}+\text{死球数}+\text{犠飛数}}$$

$$\text{長打率} = \frac{1\times\text{単打}+2\times\text{二塁打}+3\times\text{三塁打}+4\times\text{本塁打}}{\text{打数}}$$

直観的には、出塁率が高い：巧打　長打率が高い：強打

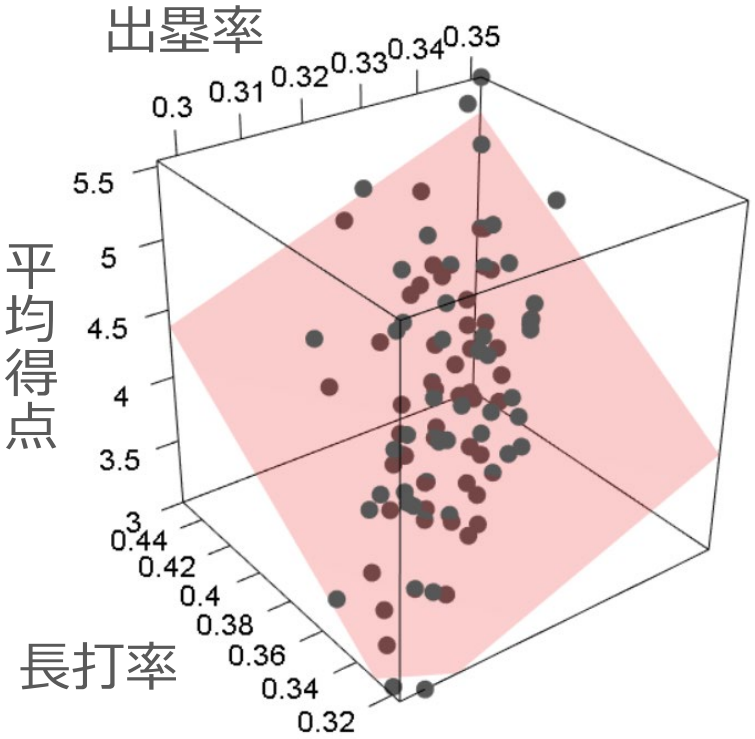
OPSを用いると、決定係数が0.8938となり、かなり1に近づく

→ OPSがどの程度だと、どれくらい点が取れるかが分かる

データの出典：パ･リーグ.com https://pacificleague.com/（2021年9月27日閲覧）

平均得点（線形重回帰分析）

2005～2020年プロ野球（パ･リーグ）チーム成績の回帰分析



素朴な疑問

出塁率と長打率を足して1つの量にしてしまうより、出塁率と長打率を、そのまま両方用いた方が、得点を上手く近似できるのでは？

重回帰分析

単回帰分析では、ある量（得点）を別の1種類の量（OPS）から計算する近似式を求める。

これに対し、別の2種類（または3種類以上）の量（出塁率と長打率）から計算する近似式を求める手法を、重回帰分析と言う。

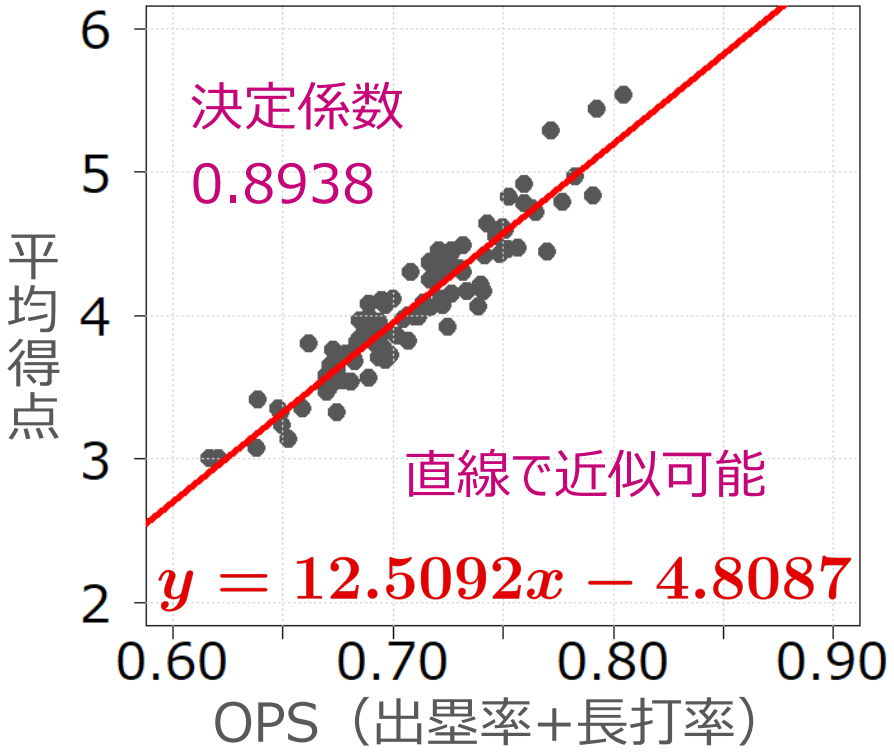
幾何学的な意味

線形の単回帰分析は、2次元（OPSと得点）の平面における点の散らばりを、直線で近似していた。線形の重回帰分析では、３次元（出塁率、長打率、得点）の空間における点の散らばりを、平面で近似する。

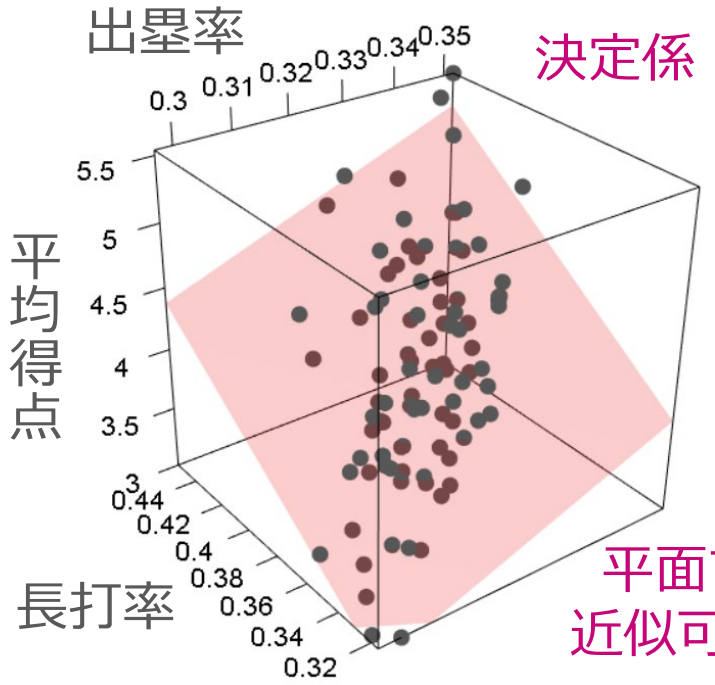
平均得点（線形重回帰分析）

2005～2020年プロ野球（パ･リーグ）チーム成績の回帰分析

単回帰分析



重回帰分析



データの出典：パ･リーグ.com
https://pacificleague.com/
（2021年9月27日閲覧）

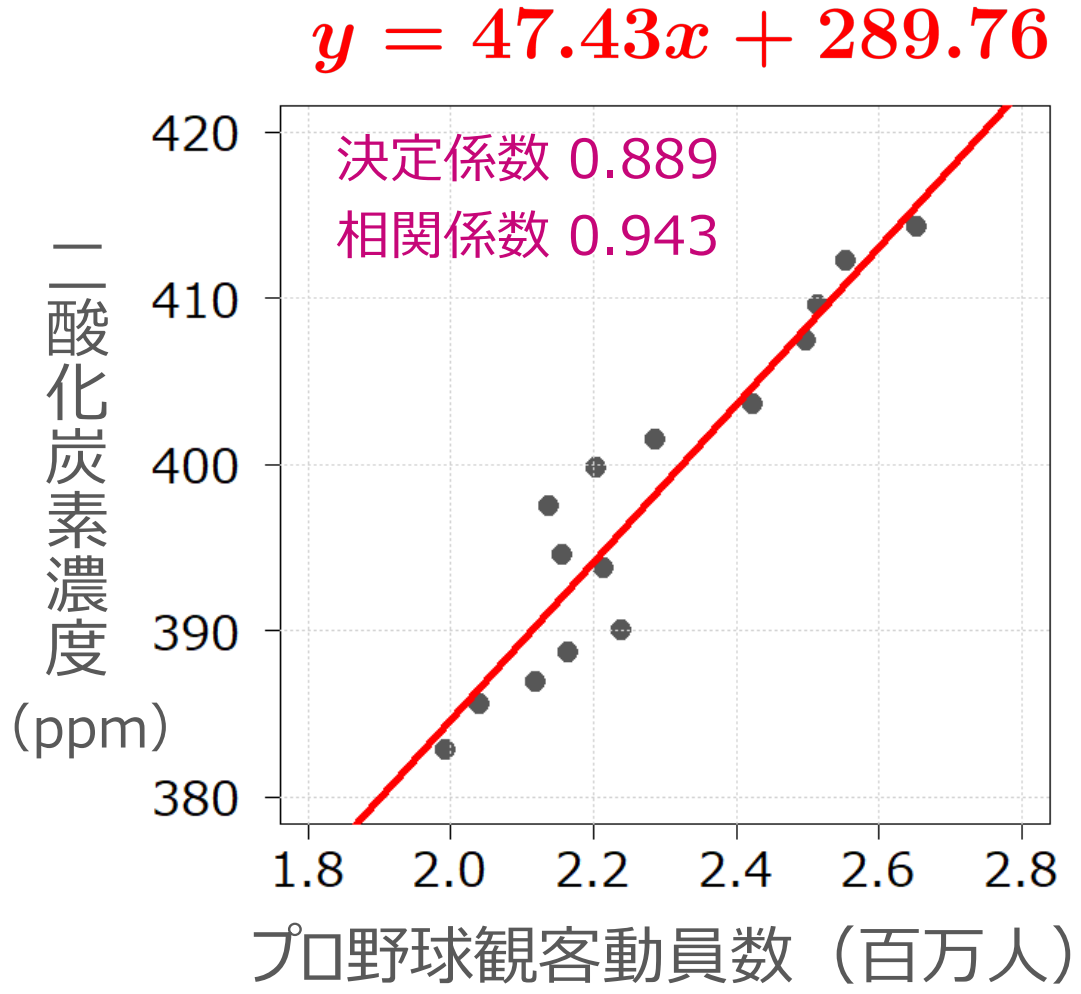
x ：出塁率
y ：長打率
z ：平均得点

$$z = 16.4230x + 11.0904y - 5.5340$$

※出塁率 ＝ （ヒット数＋四球数＋死球数）÷（打数＋四球数＋死球数＋犠飛数）
※長打率 ＝ 塁打数÷打数 （塁打：単打＝1、二塁打＝2、三塁打＝3、本塁打＝4）
※OPSより、1.5×出塁率＋長打率の方が近似精度が高い？（でも、大差なし）

平均得点（因果関係）

野球の観客が増えると、CO_2濃度が上昇？



2005年～2019年の年毎のデータを使用

※プロ野球観客動員数の公表値は、2004年までは正確な値ではないので、2005年以降を使用。また、2020年以降は、COVID-19の影響があるので除外。

縦軸：各年の日本上空のCO_2濃度（ppm）
（観測地点は、岩手県の綾里）

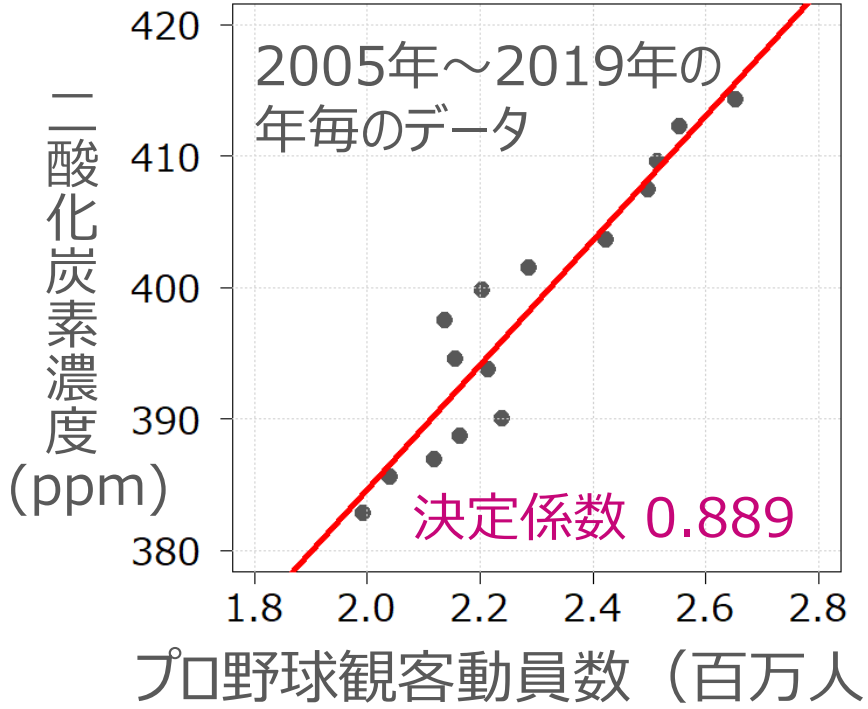
出典：気象庁『二酸化炭素濃度の観測結果』（2024.3.31閲覧）
https://www.data.jma.go.jp/ghg/kanshi/obs/co2_yearave.html

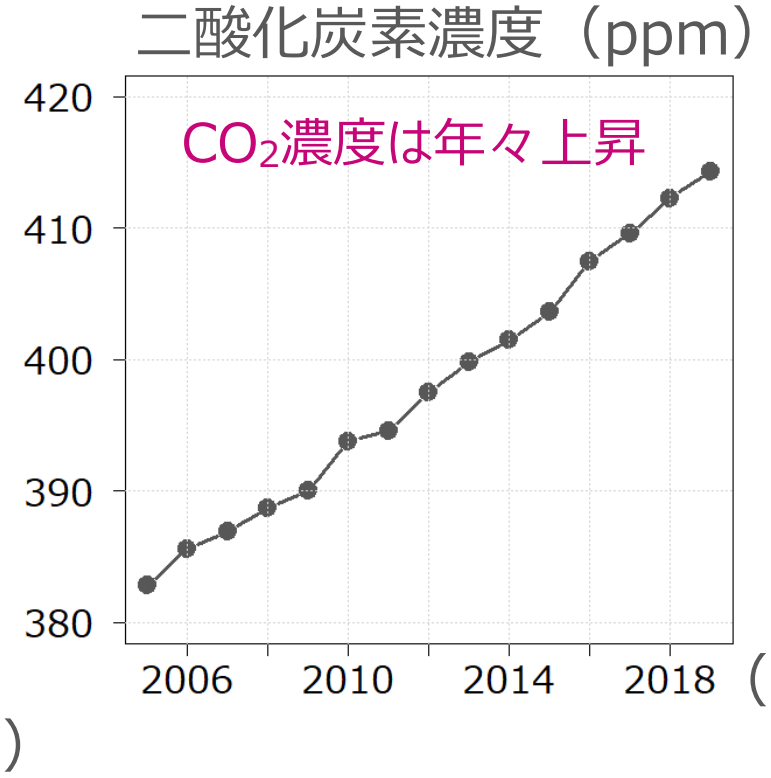
横軸：各年のプロ野球12球団合計観客動員数（百万人）

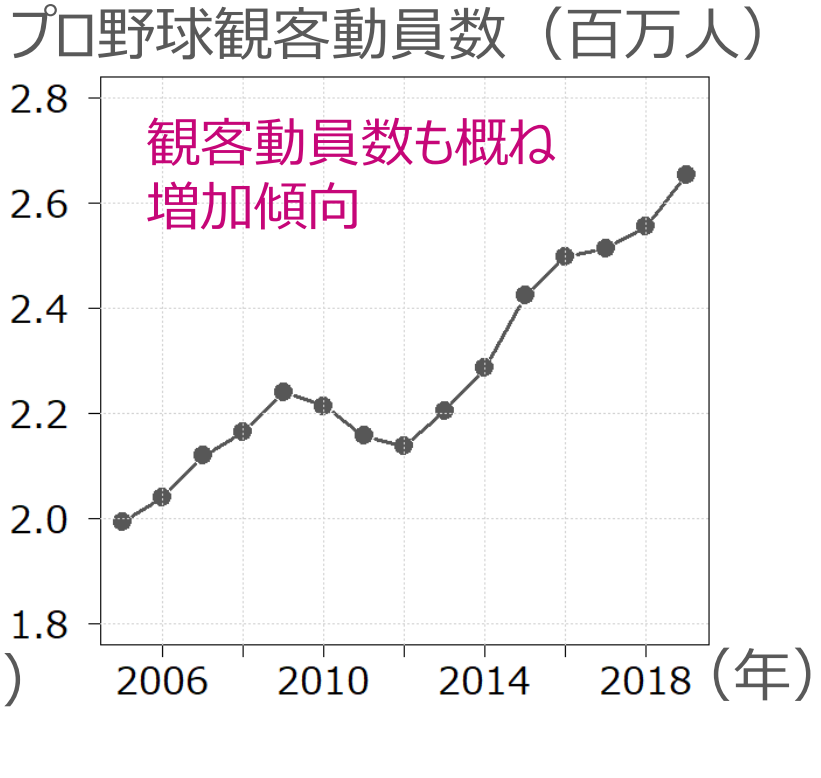
出典：日本野球機構『統計データ』（2024.3.31閲覧）
https://npb.jp/statistics/

平均得点（因果関係）

相関関係と因果関係を混同しないこと!!







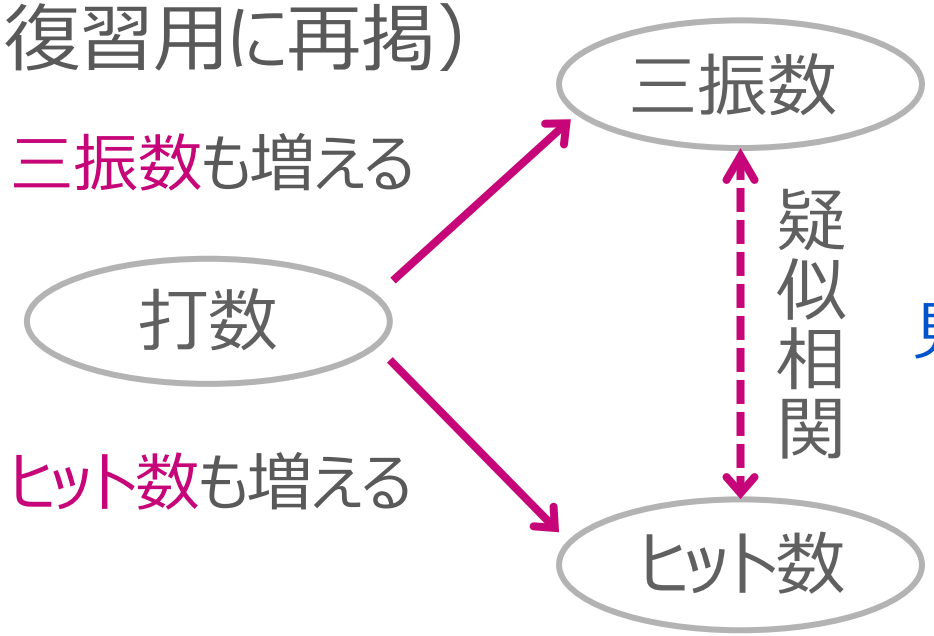
- 常識的に考えて、プロ野球の観客動員数が増えたから、岩手上空のCO_2濃度が上昇したわけではない。
- 逆に、CO_2濃度が上昇したから、観客動員数が増えたわけでもない。
- 一般に、時間の経過に伴い同じような動き方をするデータ同士は、直接の因果関係がなくても、強い関係性があるように見える（疑似相関）。

平均得点（因果関係）

疑似相関（復習用に再掲）

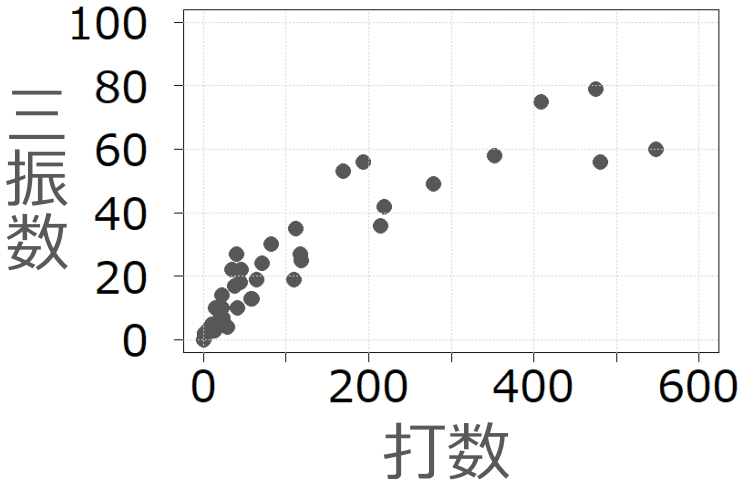
打数が増えれば、三振数も増える

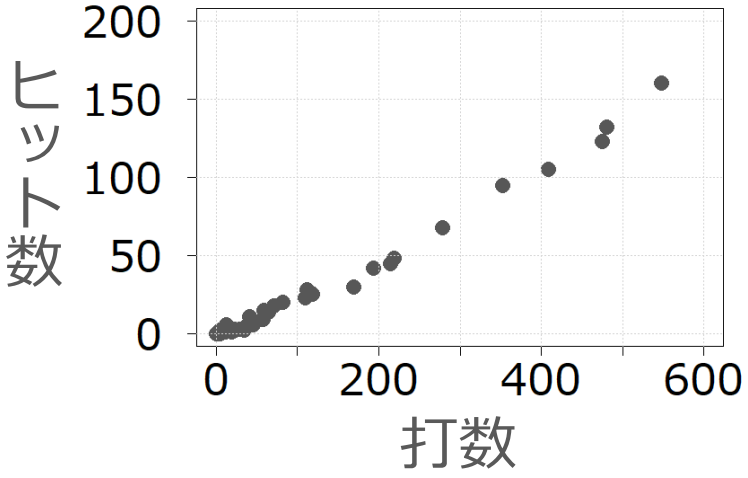
打数が増えれば、ヒット数も増える

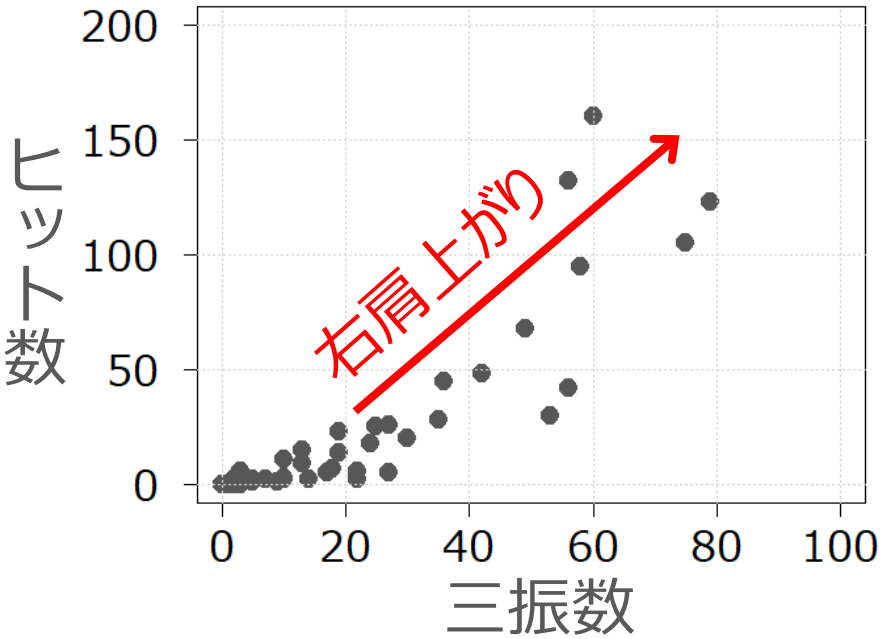


データ出典：中日ドラゴンズ公式HP
シーズン打撃成績（2022年2月12日閲覧）
https://dragons.jp/teamdata/batting.html

見せかけの関係性が生じる





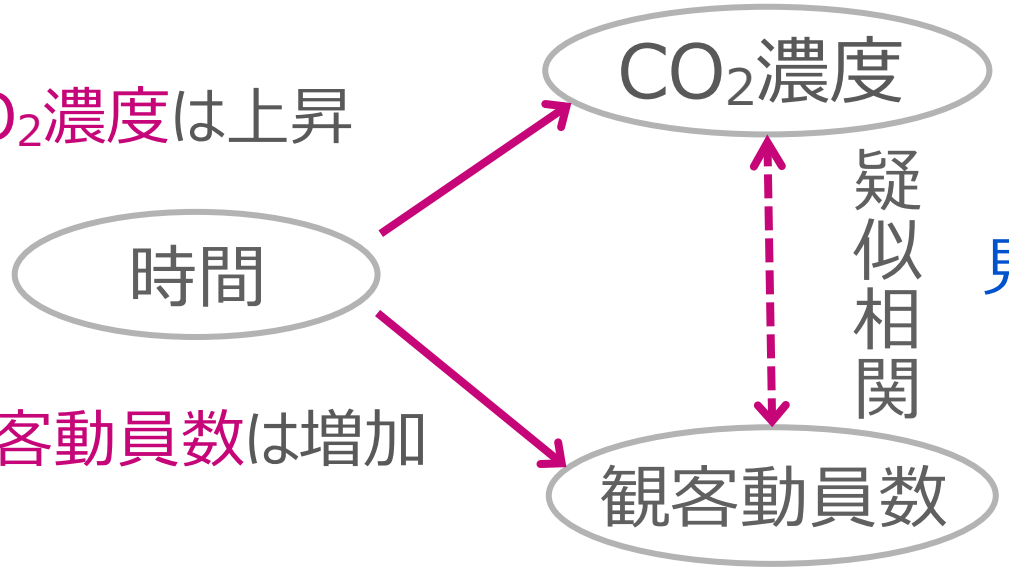


平均得点（因果関係）

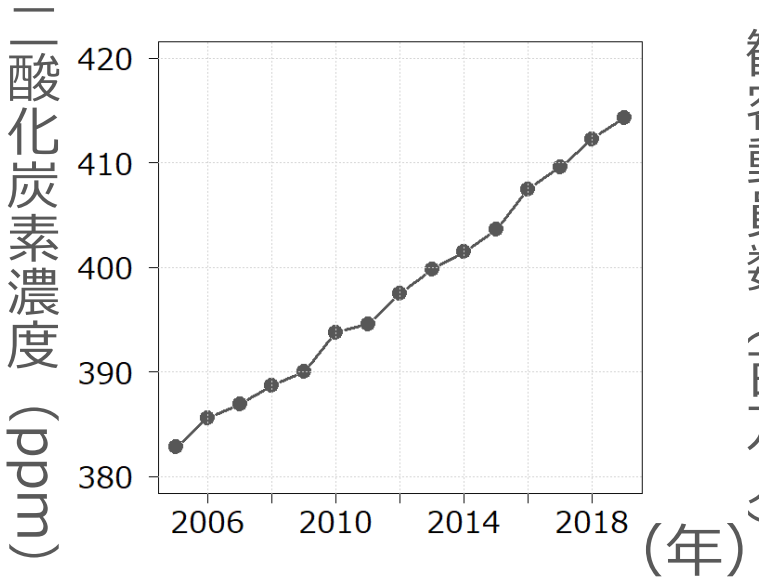
疑似相関

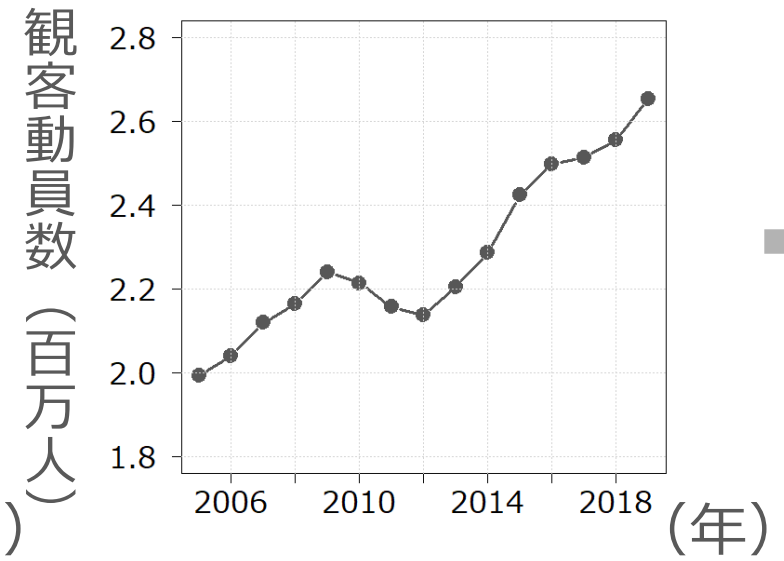
時間の経過とともに、CO_2濃度は上昇

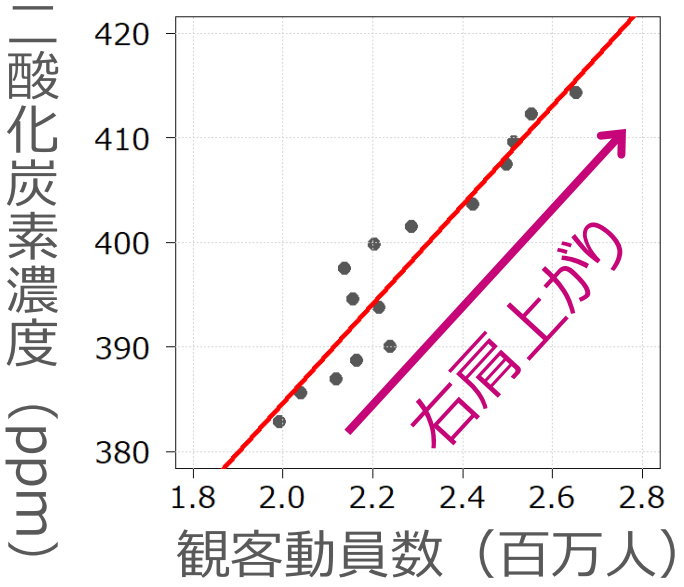
時間の経過とともに、観客動員数は増加



見せかけの関係性が生じる







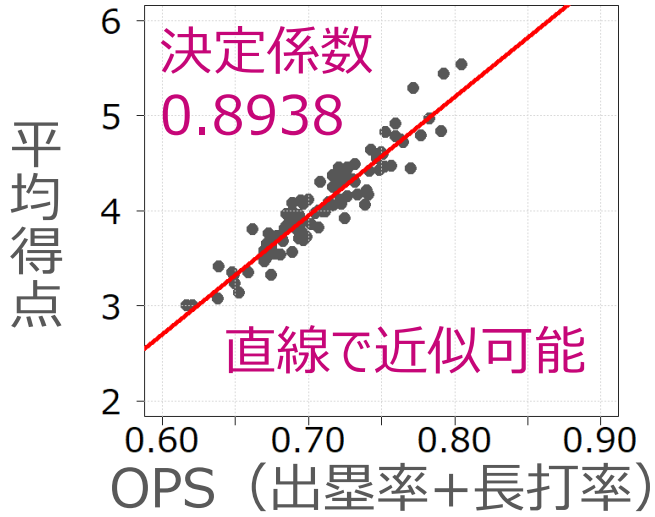
平均得点（まとめ）

試合の得点（結論）

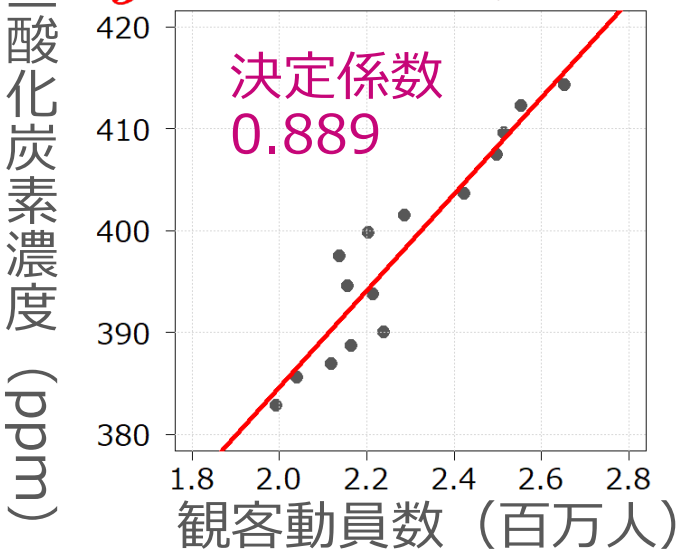
- チームの1試合当たりの平均得点は、OPS（出塁率＋長打率）と、ほぼ直線的な関係にある。
- 数式で表すと、平均得点が、OPSの1次式で近似できる。
- このように近似することを、（線形の）回帰分析という。
- 一般に、2種類のデータが、互いに直線的な関係にあるからと言って、因果関係があるとは限らないことに注意が必要（観客動員数とCO_2濃度の例参照）。
- ただし、野球のルールを考えれば、出塁率や長打率が増加すれば、得点が入りやすくなると考えられる。

結論：OPS（出塁率＋長打率）はチームの得点力を表す精度の良い指標の1つ

$$y = 12.5092x - 4.8087$$



$$y = 47.43x + 289.76$$

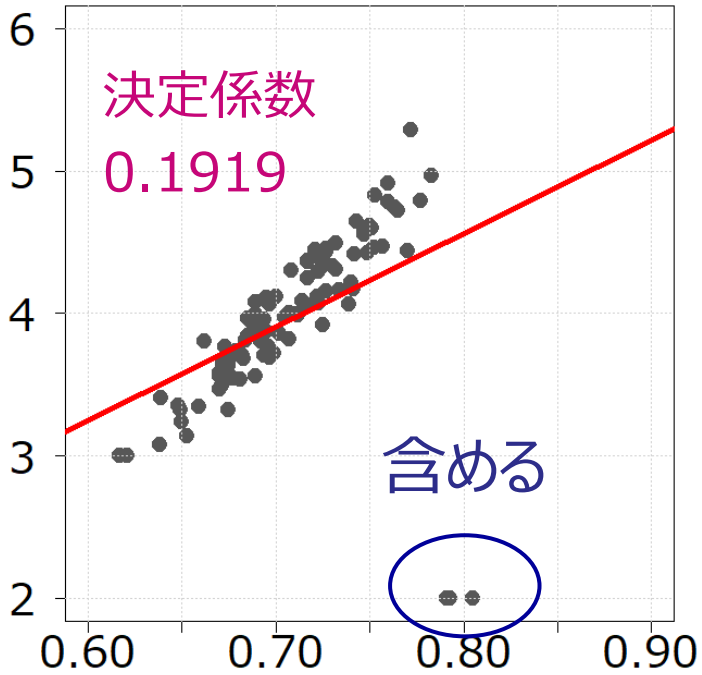


平均得点（補足）

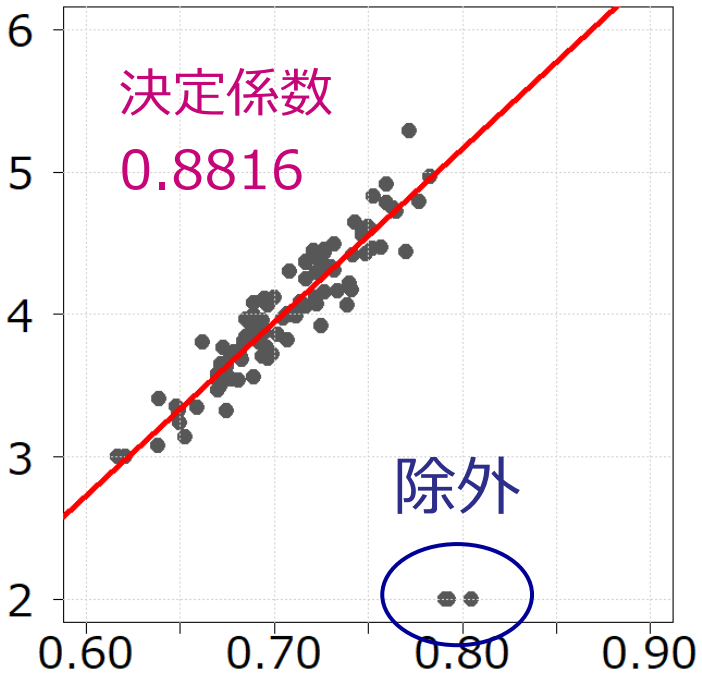
回帰分析の留意事項

データの出典：パ・リーグ.com（https://pacificleague.com/）のパリーグのチーム平均得点、OPSのデータ（前掲）を、説明用に改編した架空データ

データをよく見て解析しないと、的外れな解析になってしまう。
いきなり回帰分析するのではなく、まずは散布図を描き、状況確認することが大切。



外れ値に引きずられる



外れ値以外を上手く近似

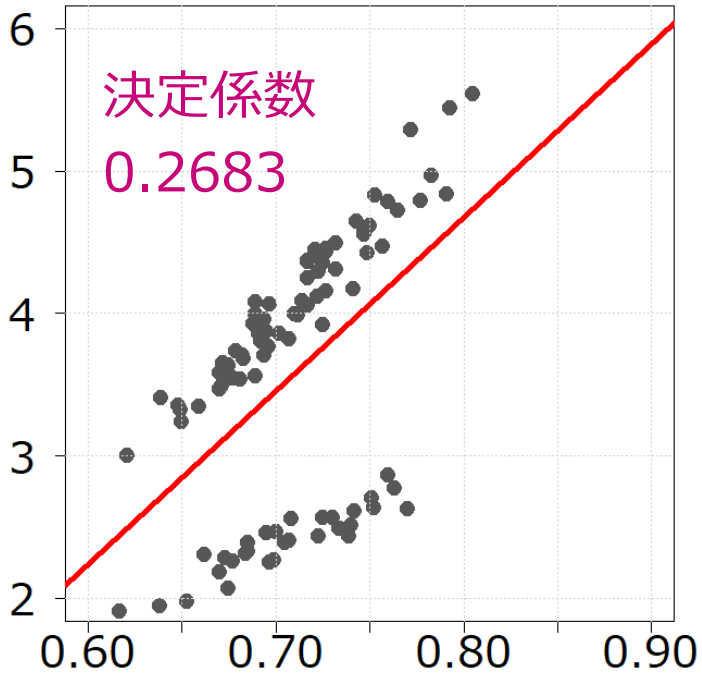
- 解析対象のデータの中に、外れ値（他とはかけ離れた値）が含まれているときは要注意。
- そのまま解析してしまうと、外れ値に引きずられて、解析結果が大きく歪められる。
- 外れ値を除外することで、データの持つ本来の性質や法則性を的確に抽出できるようになる。

平均得点（補足）

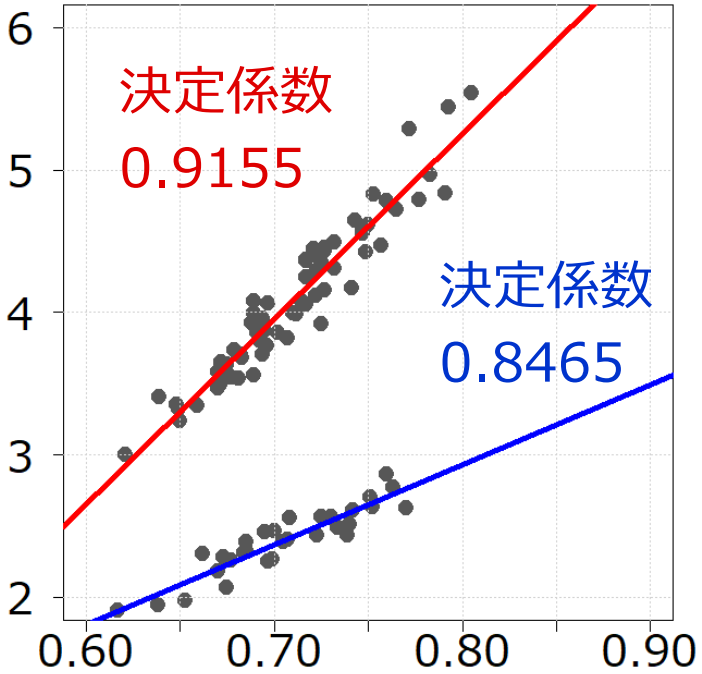
回帰分析の留意事項

データの出典：パ･リーグ.com（https://pacificleague.com/）のパリーグのチーム平均得点、OPSのデータ（前掲）を、説明用に改編した架空データ

データをよく見て解析しないと、的外れな解析になってしまう。
いきなり回帰分析するのではなく、まずは散布図を描き、状況確認することが大切。



2群を中途半端に近似



群ごとに別々に近似

- 解析対象のデータが、2つのグループに分かれているときは要注意（例えば、プロ野球のデータと少年野球のデータ）。
- まとめて解析してしまうと、どっちつかずの中途半端な結果になる。
- グループごとに、別々に解析することで、それぞれのグループに対し、データの持つ本来の性質や法則性を的確に抽出できるようになる。

https://www.math.sci.ehime-u.ac.jp/~matsuura/

打席数のデータに隠された数理曲線

非線形回帰分析

～ データを曲線で近似 ～

打席数（使用データ）

打席数のデータ

打席数とは、打席（バッターボックス）に立った回数。

ここでは、投手（ピッチャー）も含め、2021年に、中日ドラゴンズに所属した各選手の2021年シーズンにおける一軍公式戦での打席数のデータを分析する。

分析対象のデータ

背番号	選手	打席数
0	高松	123
00	石岡	11
1	京田	448
3	高橋周	520
以下略		

- 打席数が10以上の37選手を扱う
- 最大値：596（大島選手）、最小値10（山下選手）
- データ出典：中日ドラゴンズ公式サイト シーズン打撃成績（2022年2月12日閲覧）から算出
  ※現時点では、直近のシーズンのデータに更新済
  https://dragons.jp/teamdata/batting.html

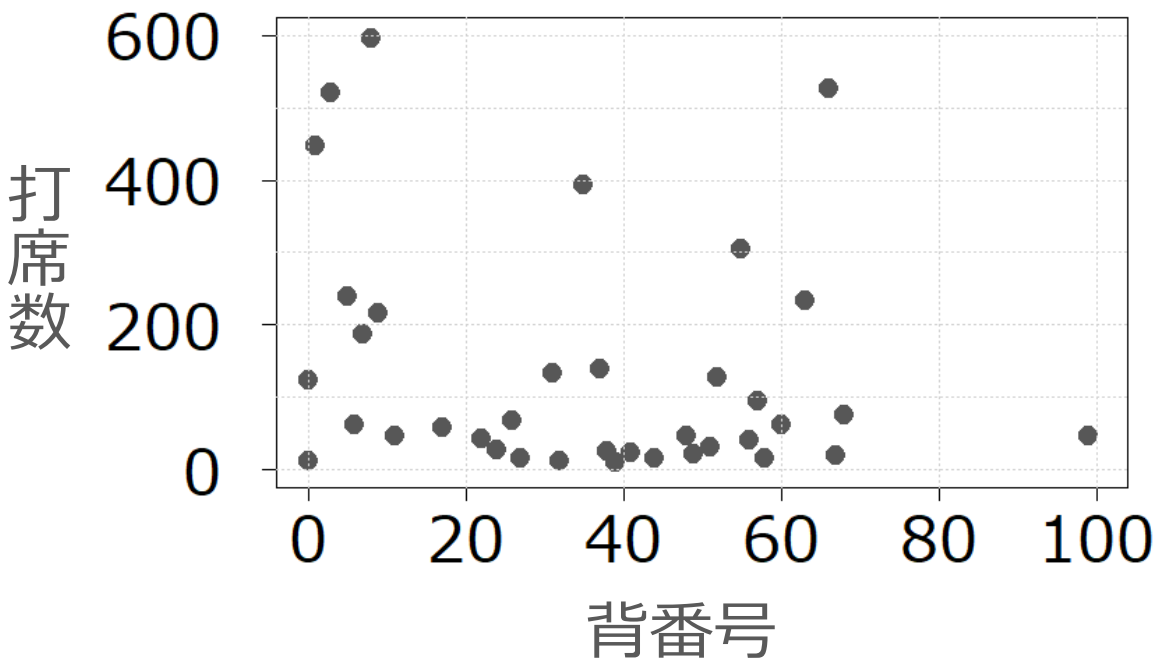
打席数（散布図）

打席数の可視化（図を描く）

2021年、打席数10以上のドラゴンズ所属選手

まずは、横軸を背番号、縦軸を打席数とする散布図を描く。

打席数の散布図



※石岡選手の背番号00は便宜上0とする。

図から何が分かる？

- 打席数200未満の選手が多い。
- 背番号と打席数は、あまり関係なさそう。
  （相関係数は約 −0.244）
- この図だと、明確な法則性が見つけにくい。

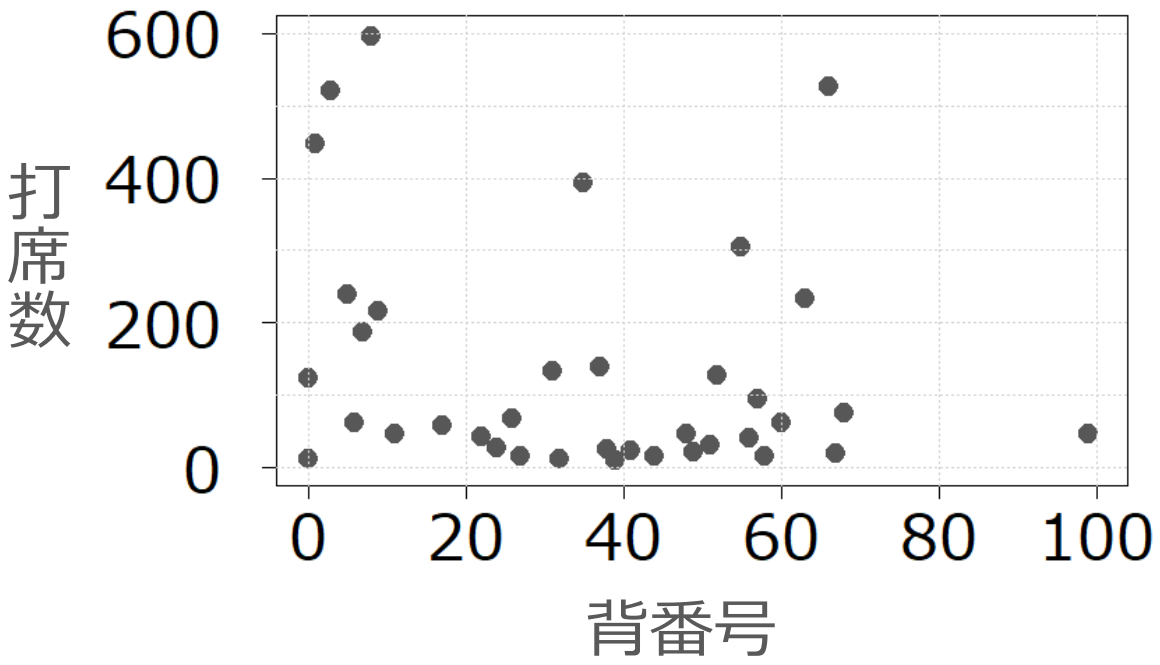
図の描き方を工夫しよう！

打席数（並べ替え）

打席数の可視化（図を改善） 2021年、打席数10以上のドラゴンズ所属選手

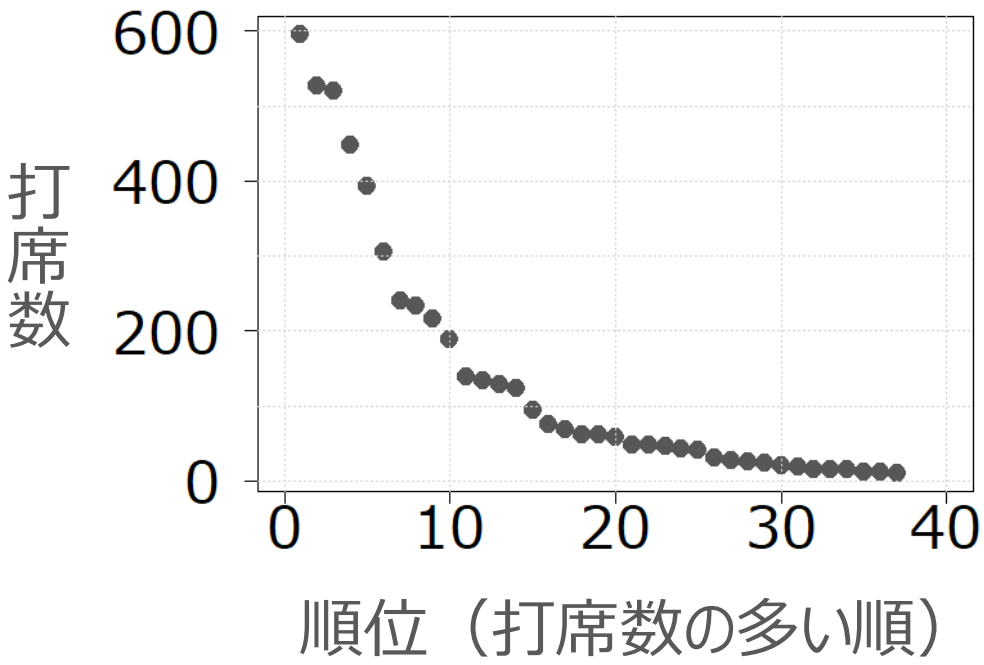
横軸を「背番号」から「順位（打席数の多い順）」に変更

打席数の散布図



規則性が良く分からない

打席数（打席数の多い順）



規則性がありそう

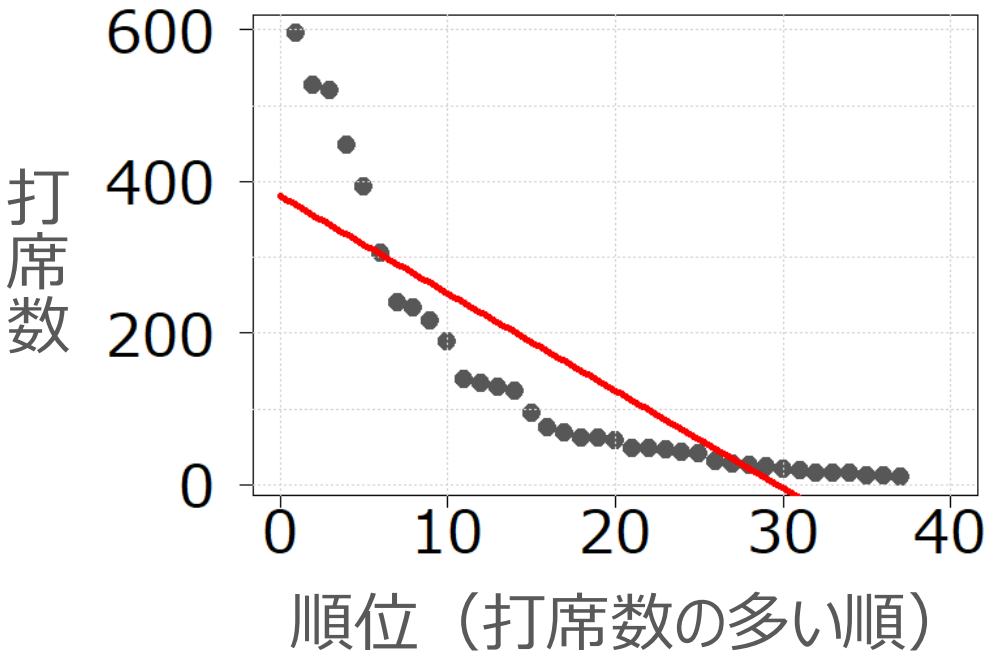
打席数（線形回帰分析）

打席数を直線で近似

2021年、打席数10以上のドラゴンズ所属選手

とりあえず、打席数を直線で近似（線形回帰分析）

$$y = -12.858x + 380.94$$



線形回帰してみた結果

- 誤差が大きくなる。
- 直線で近似するには無理がある。
- でも、回帰分析をしたい。

データを加工して、直線っぽくしよう！

打席数（対数）

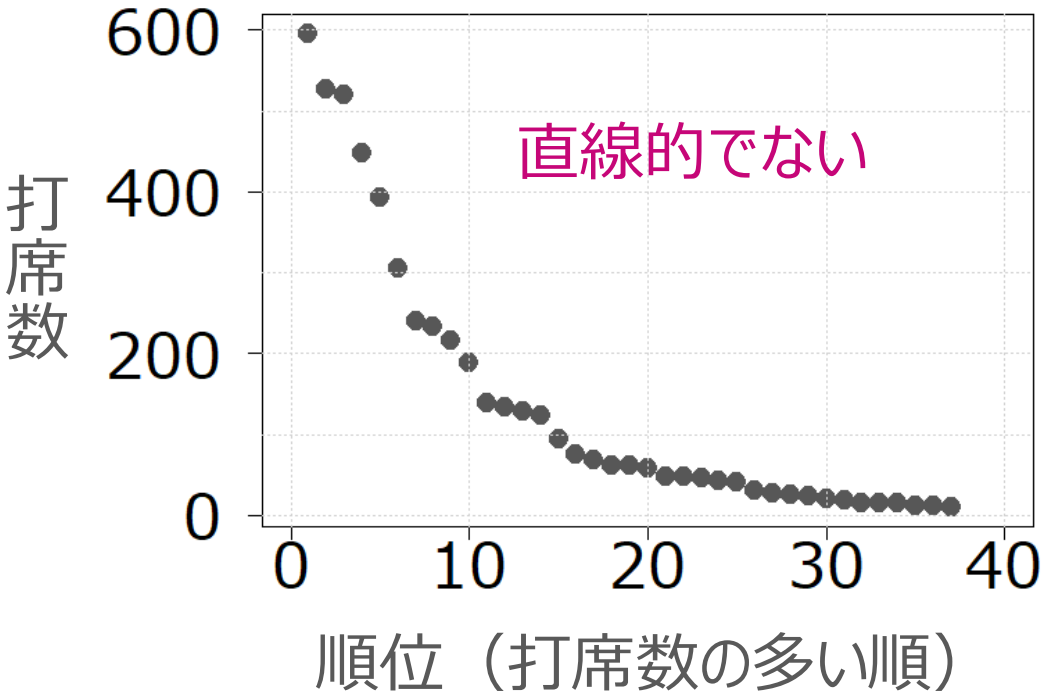
打席数のデータを加工

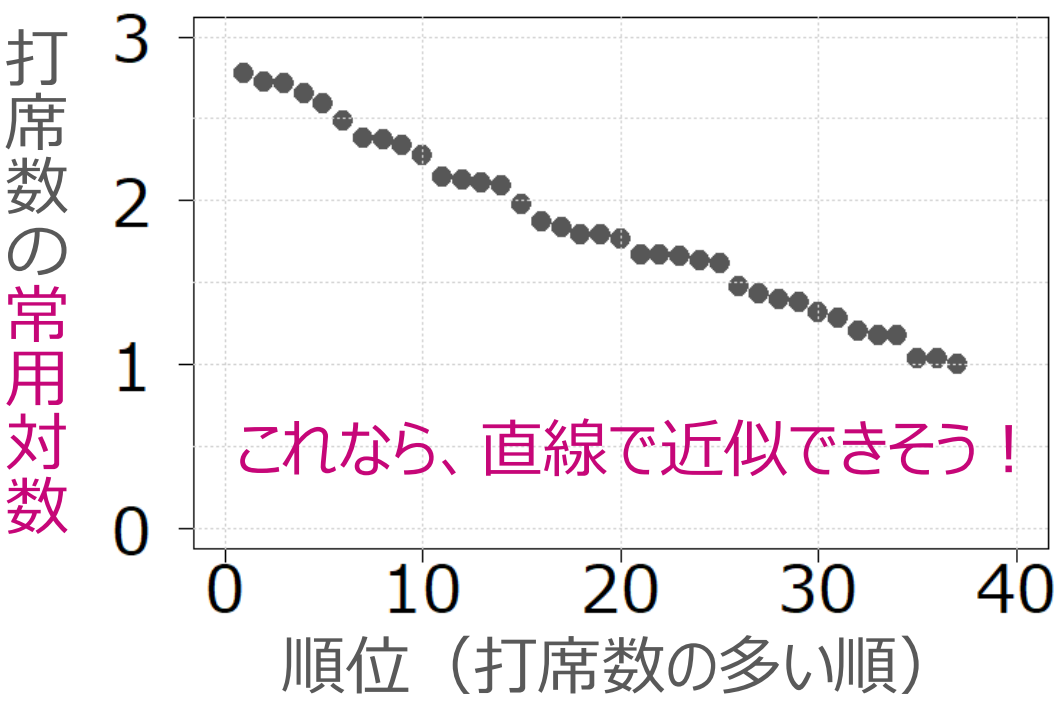
2021年、打席数10以上のドラゴンズ所属選手

打席数の常用対数をとってみる。

つまり、打席数 m のとき、縦軸の値を、m そのものではなく、$\log_{10} m$ に変える。

※10を何乗すると m になるかを表したのが、常用対数 $\log_{10} m$





打席数（対数）

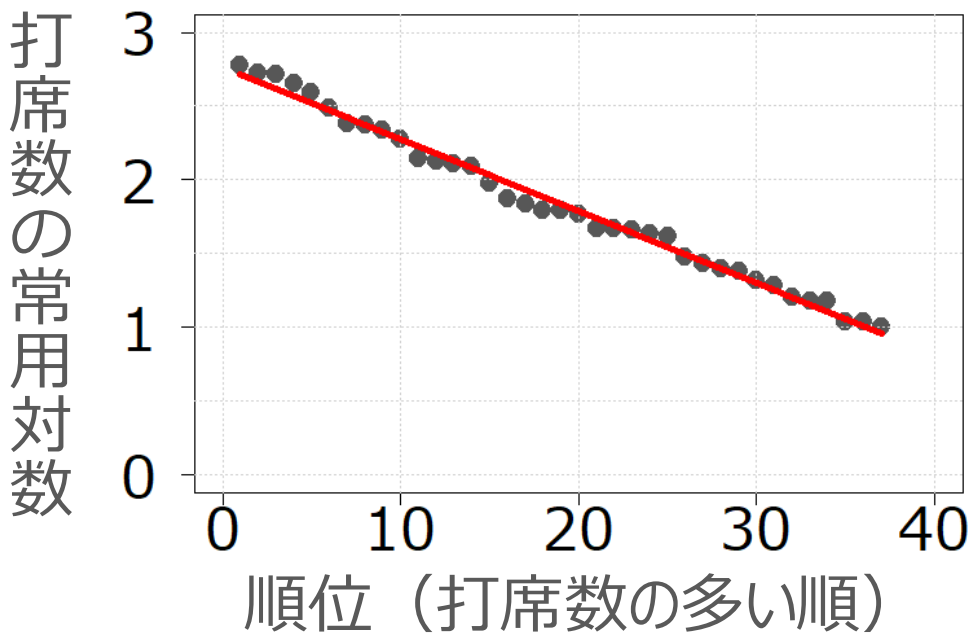
打席数の対数を直線で近似

2021年、打席数10以上のドラゴンズ所属選手

打席数の対数を直線で近似してみる。

打席数（常用対数）

$$y = -0.0488x + 2.7639$$



対数をとった上で、線形回帰してみた結果

打席数の常用対数をとった値は、
直線でかなり上手く近似できていそう。

縦軸を対数を取る前に戻したらどうなる？

打席数（非線形回帰分析）

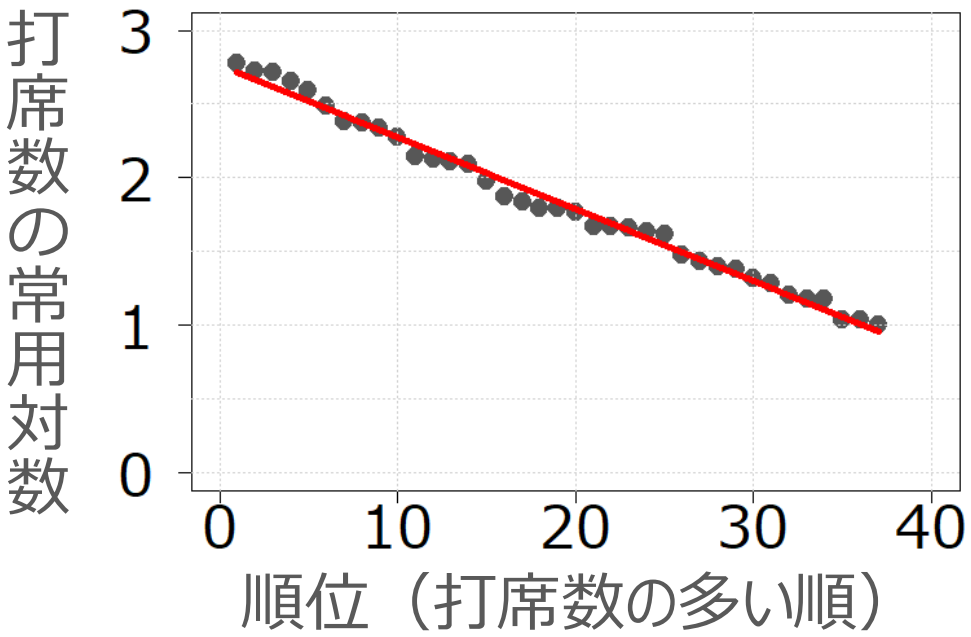
打席数のデータを元に戻す

2021年、打席数10以上のドラゴンズ所属選手

対数関数の逆関数は指数関数なので、縦軸を元に戻すと、近似直線は近似曲線（指数関数）に変換される。
※xをyに移す関数に対し、逆にyをxに移す関数（つまり、元に戻す関数）が逆関数。

打席数（常用対数）

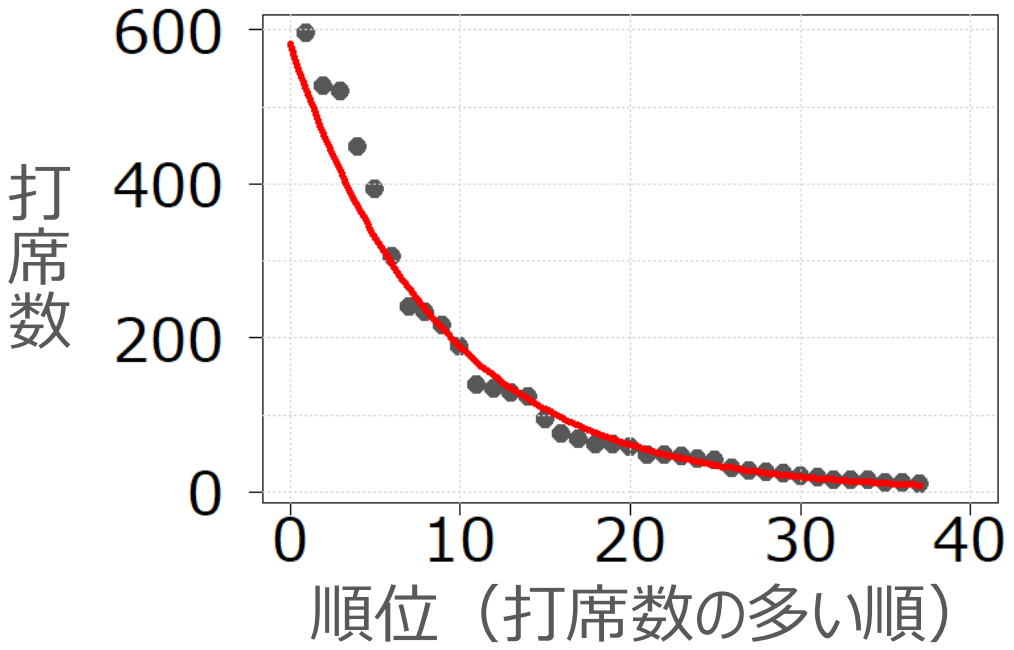
$$y = -0.0488x + 2.7639$$



縦軸を元に戻す

打席数（指数関数で近似）

$$y = 580.69 \times 10^{-0.0488x}$$



打席数（非線形回帰分析）

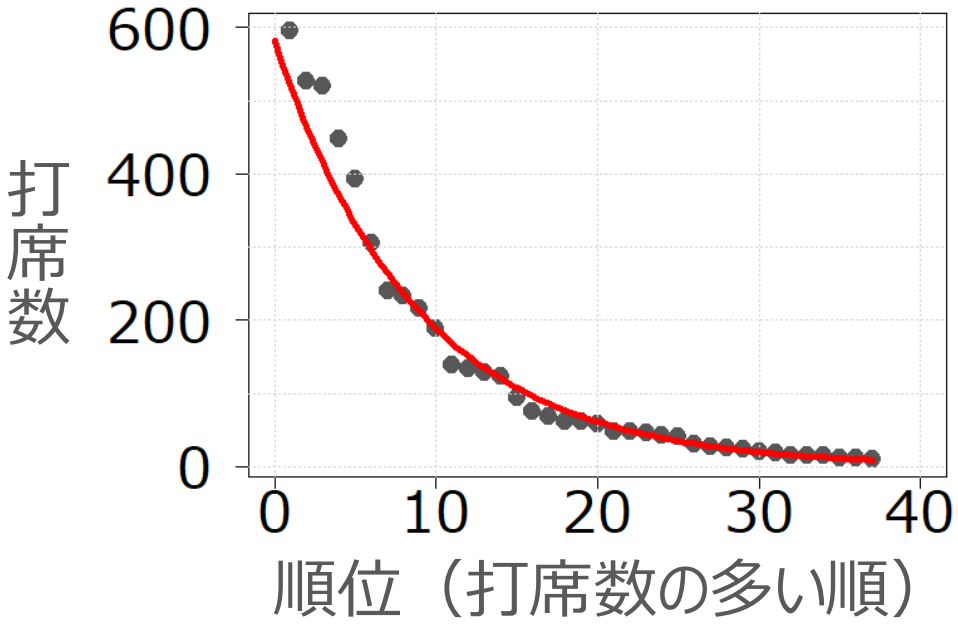
近似曲線の改良

2021年、打席数10以上のドラゴンズ所属選手

打席数は指数関数で比較的良く近似できている。でも、順位の小さなところで誤差が大きい。
そこで、常用対数への加工は止め、打席数そのものを、直接、指数関数で近似する。

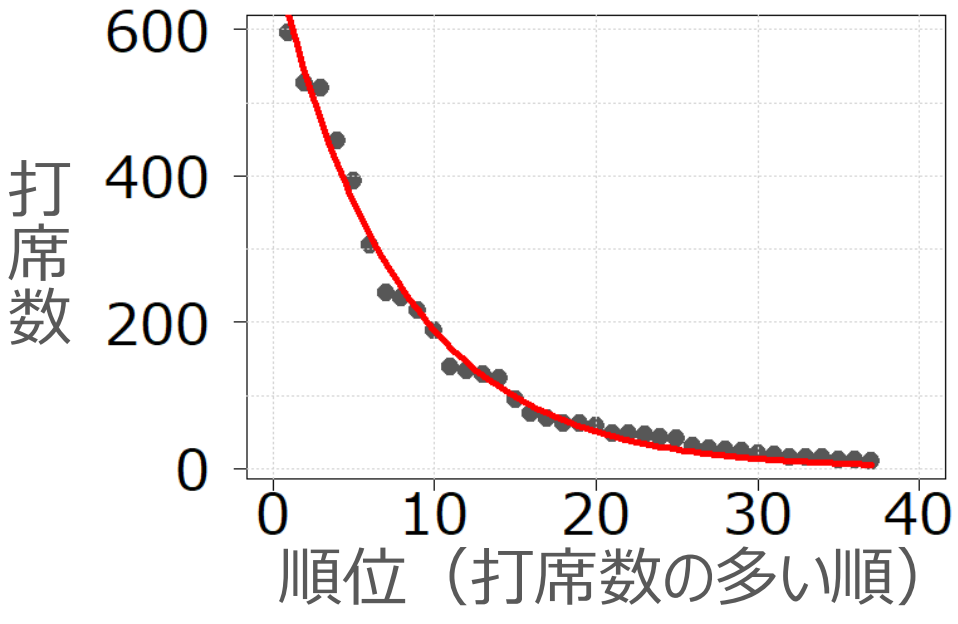
対数を取って近似し、元に戻したもの

$$y = 580.69 \times 10^{-0.0488x}$$



対数を経ずに直接、指数関数で近似

$$y = 704.91 \times 10^{-0.0572x}$$



順位が小さい所も、上手く近似できている！

打席数（非線形回帰分析）

打席数のデータ（改めて）

2021年シーズンに、中日ドラゴンズに所属した各選手について、
2021年シーズンにおける打席数のデータを分析。

分析対象のデータ

背番号	選手	打席数
0	高松	123
00	石岡	11
1	京田	448
3	高橋周	520
以下略		

打席数が10以上の37選手を扱った。

10以上とした理由：

実は、対数（log）を用いたかったから（恣意的）。

打席数0だと対数がとれない（$\log_{10} 0$ は存在しない）。

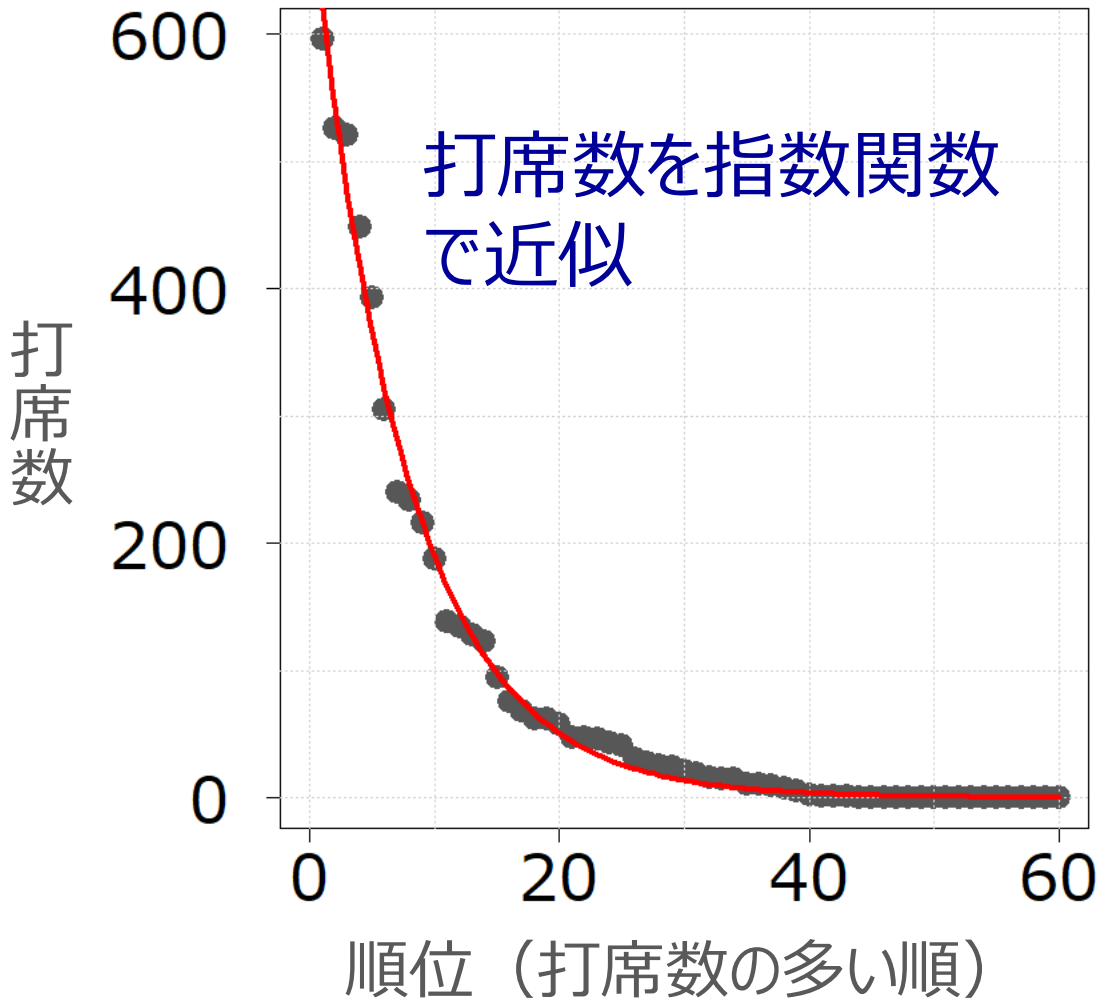
また、打席数が小さいと、誤差が強調されてしまう。

でも、恣意的にデータを取り扱うのは避けるべき。

打席数10未満のデータも使い、対数を用いず直接、指数関数で近似してみたら？

打席数（まとめ）

打席数のデータの分析（結論）



2021年、ドラゴンズ所属選手のデータ
（一軍の試合に出場した全60選手分）

打席数を多い順に並べたら、指数関数で近似できることが分かった。

$$y = 705 \times 10^{-0.0572x}$$

一般に、データを曲線で近似して分析を行うことを、「非線形回帰分析」と呼ぶ。

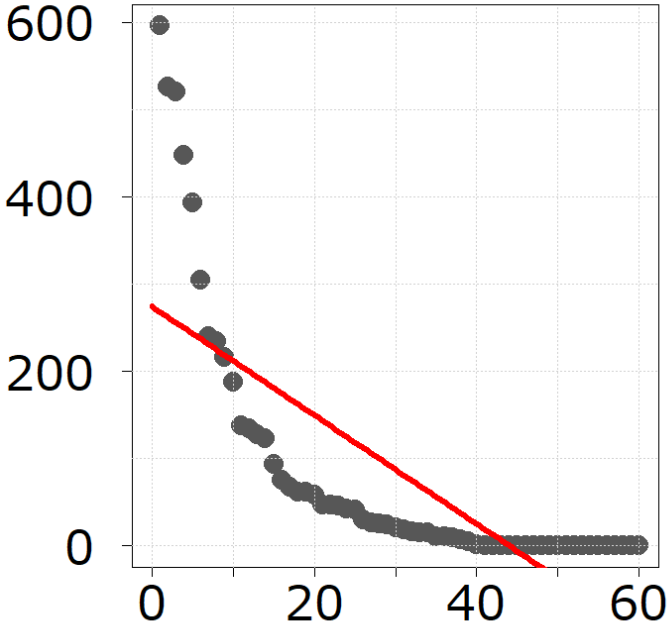
データ出典：中日ドラゴンズ公式サイト
シーズン打撃成績（2022年2月12日閲覧）
※現時点では、直近のシーズンのデータに更新済
https://dragons.jp/teamdata/batting.html

打席数（補足）

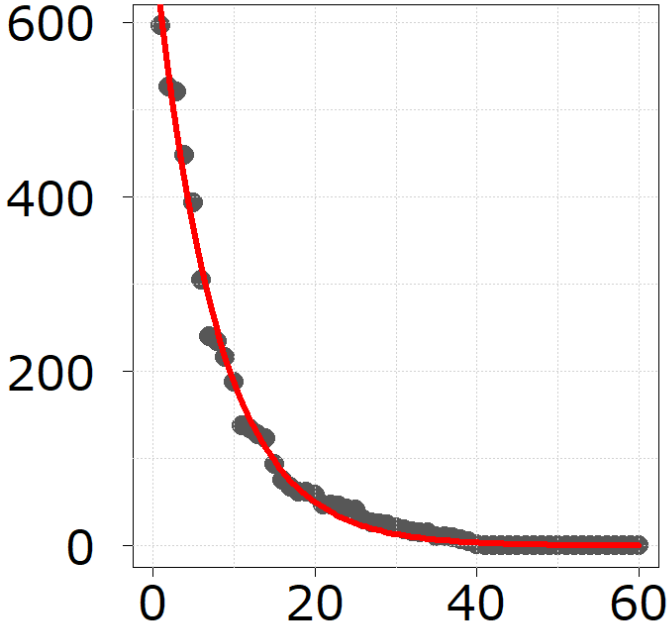
回帰分析の留意事項

データの出典：中日ドラゴンズ公式サイト シーズン打撃成績（2022年2月12日閲覧）

データをよく見て解析しないと、的外れな解析になってしまう。
いきなり回帰分析するのではなく、まずは散布図を描き、状況確認することが大切。



無理やり直線で近似しようとしても、上手く近似できない



曲線を用いれば、精度よく近似可能

- 解析対象のデータが、曲線的に分布しているとき、無理に直線で近似しようとしても、上手く近似できない。
- 回帰分析を始める前に、まずは散布図を描き、点の分布が、直線状か曲線状かを確認すると良い。

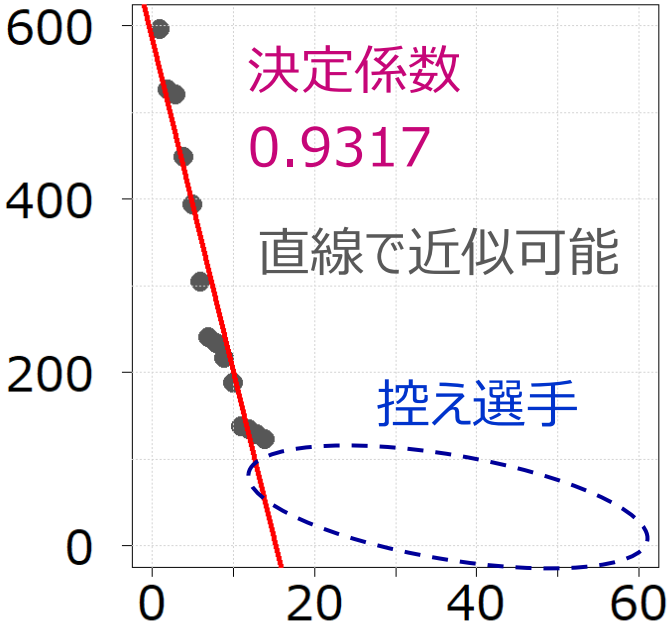
直線状 → 線形回帰
曲線状 → 非線形回帰

打席数（補足）

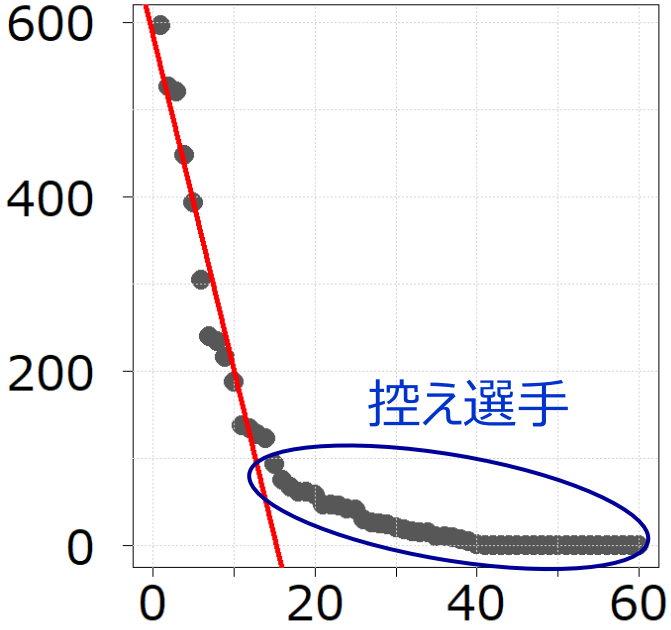
回帰分析の留意事項

データの出典：中日ドラゴンズ公式サイト シーズン打撃成績（2022年2月12日閲覧）

分析に用いるデータが偏っていると、分析結果も偏ってしまう。
偏ったデータによる分析結果を、無理に一般化し、拡大解釈しないよう留意が必要。



レギュラー・準レギュラー選手のみのデータで線形回帰分析



得られた近似直線は控え選手のデータには不適合

- レギュラー・準レギュラーの選手だけを対象にすれば、打席数のデータは、直線で精度よく近似可能。
- しかし、レギュラー・準レギュラーの選手だけのデータから、控え選手について推測することはできない。
- 控え選手も含めた知見を得たければ、分析対象のデータに控え選手も含めないといけない。

選手それぞれの特徴は？

主成分分析

～ 選手の成績を平面に図示 ～

選手の特徴（使用データ）

打撃（バッティング）成績のデータ

2021年に中日ドラゴンズに所属した各選手（打数1以上の43選手）の、2021年シーズンの一軍公式戦打撃データ（出場試合、打席数、打数、得点、安打、二塁打、三塁打、本塁打、塁打、打点、盗塁、盗塁刺、犠打、犠飛、四球、故意四球、死球、三振、併殺打、打率、長打率、出塁率）を分析。

データ出典：中日ドラゴンズ公式HP
シーズン打撃成績（2022年2月12日閲覧）
※現時点では、直近のシーズンのデータに更新済
https://dragons.jp/teamdata/batting.html

分析対象のデータ

背番号	選手	試合	打席数	打数	得点	安打	二塁打	三塁打	本塁打	塁打	…	長打率	出塁率
0	高松	78	123	112	18	28	1	0	0	29	…	0.259	0.274
00	石岡	6	11	11	0	1	0	0	0	1	…	0.091	0.091
1	京田	113	448	409	41	105	7	4	3	129	…	0.315	0.302
以下略													

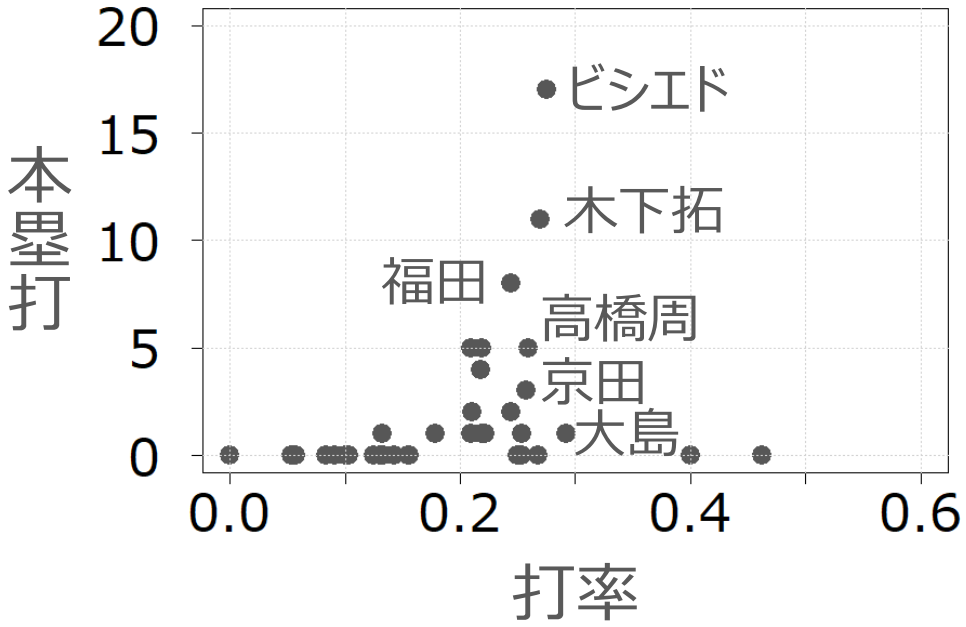
選手の特徴（散布図）

打撃データを図示

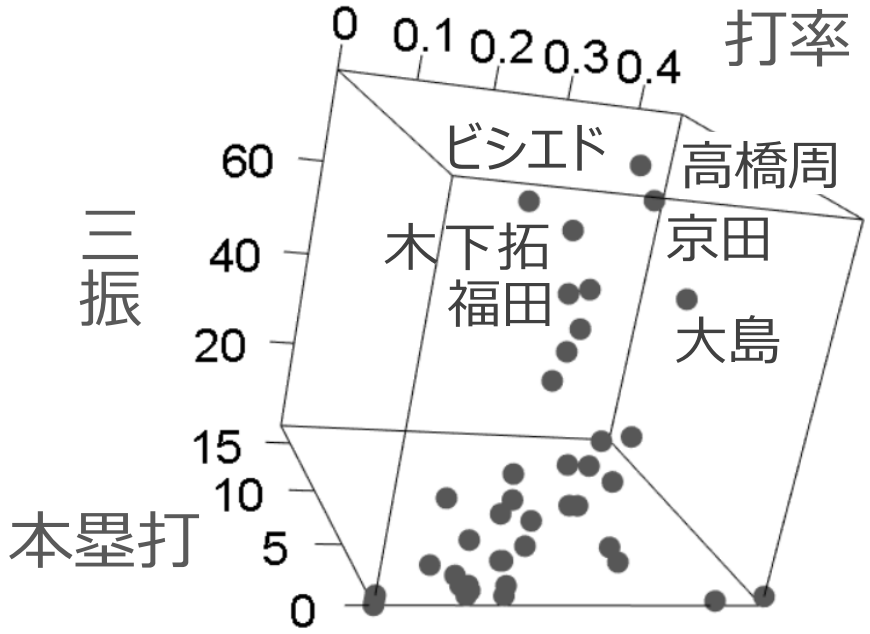
・2項目（例えば、打率と本塁打）だけなら、2次元平面に図示可能。

・3項目（例えば、打率と本塁打と三振）だけなら、（強引に）3次元空間に図示可能。

打率と本塁打のみ図示



打率と本塁打と三振のみ図示



選手の特徴（主成分分析の原理）

多くの項目のデータを図示

・3次元空間に図示すると、あまり見やすくない。見やすくできないか？

・さらに、全22項目を見やすく図示できないか？

全22項目を図示したい

⇩

「22次元空間内の点」として各打者を表現

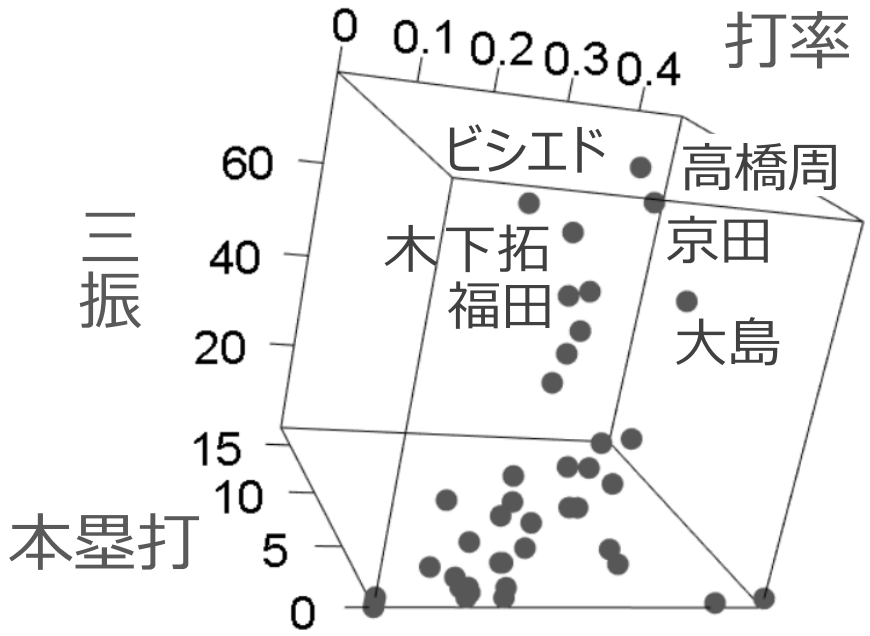
⇩

でも、「22次元空間」は人間の目では理解不能

⇩

2次元に圧縮して図示

打率と本塁打と三振のみ図示

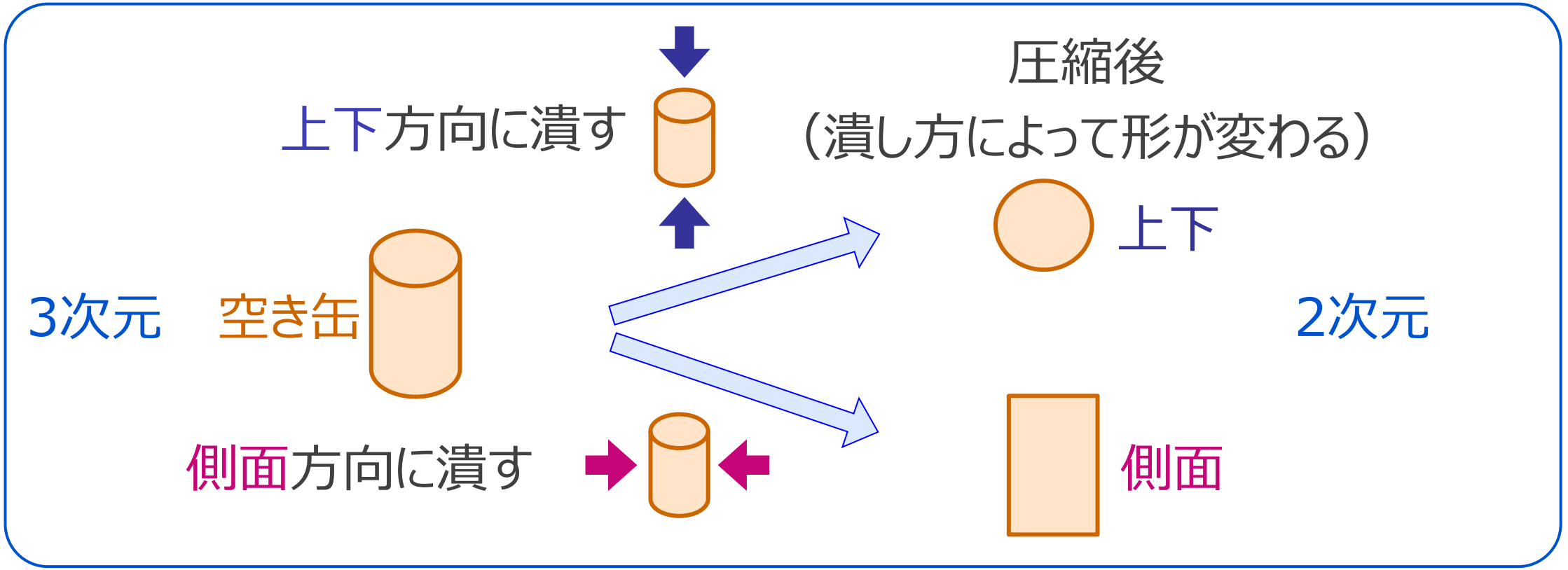


選手の特徴（主成分分析の原理）

空き缶潰し（次元を圧縮）

・3次元の立体を潰して、2次元の図形に圧縮。

・どちらの方向から潰したかで、圧縮後の形が変わる。



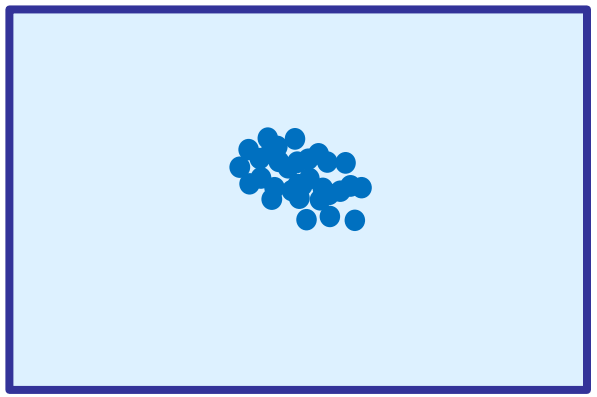
選手の特徴（主成分分析の原理）

データの次元を圧縮

・3次元のデータを潰して、2次元のデータに圧縮。
・どちらの方向から潰したかで、データの見やすさが変わる。

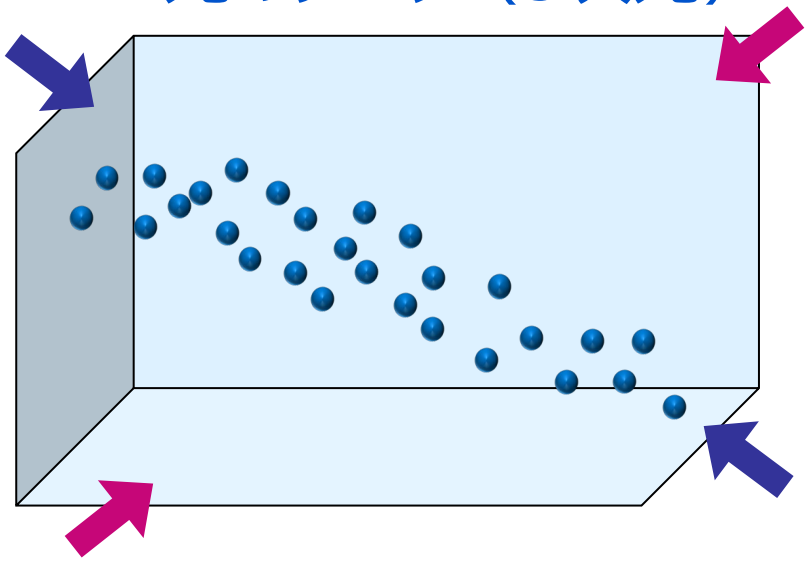
好ましくない潰し方

紺の矢印の方向に潰すと点は密集
（区別が付かなくなる）



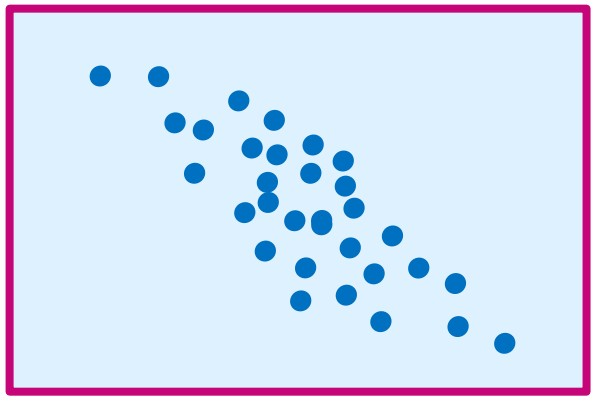
次元圧縮後（2次元）

元のデータ（3次元）



好ましい潰し方

赤の矢印の方向に潰すと点は分散
（区別が付きやすい）



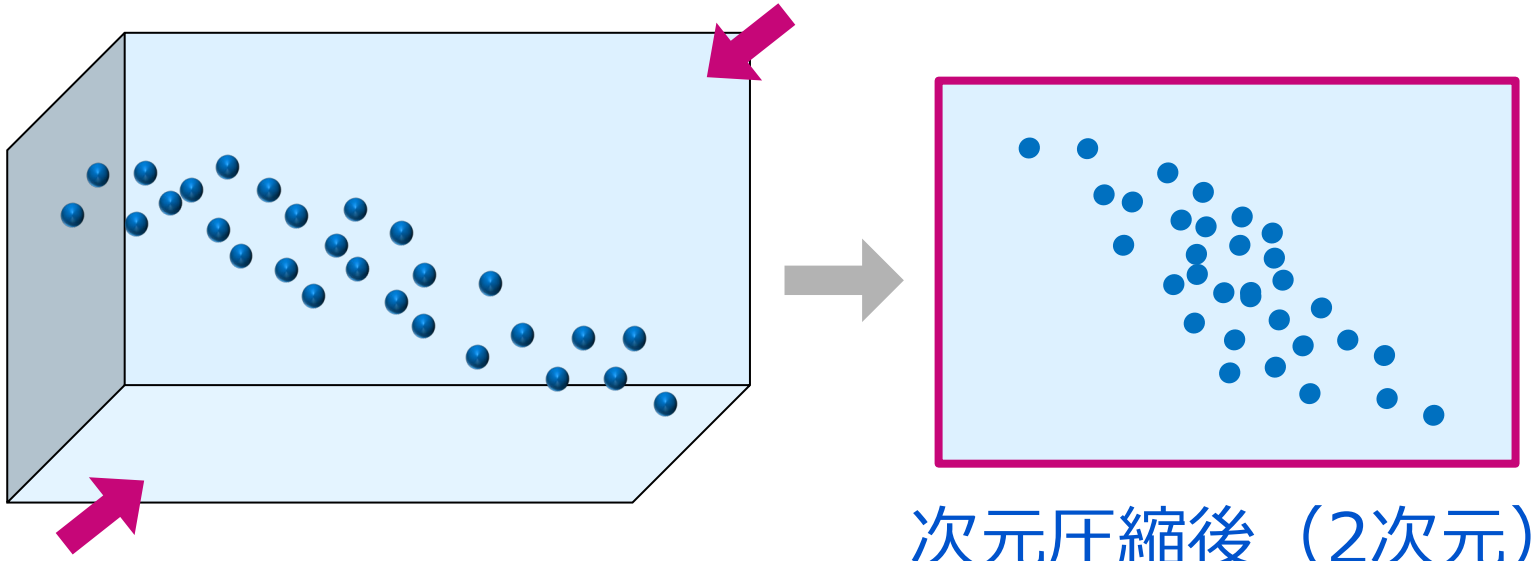
次元圧縮後（2次元）

選手の特徴（主成分分析の原理）

主成分分析

・次元の高い（項目数の多い）データの次元について、空き缶潰しと同様にして、
　次元を圧縮し、低次元（典型的には2次元）のデータに落とす。
・その際、なるべく元のデータの特徴（点の散らばり具合）を維持する方向から圧縮。
・低次元のデータにすることで、図示などにより、直観的に理解できるようになる。
・このような手法を、「主成分分析」と呼ぶ。

なるべく元のデータの
散らばりを維持



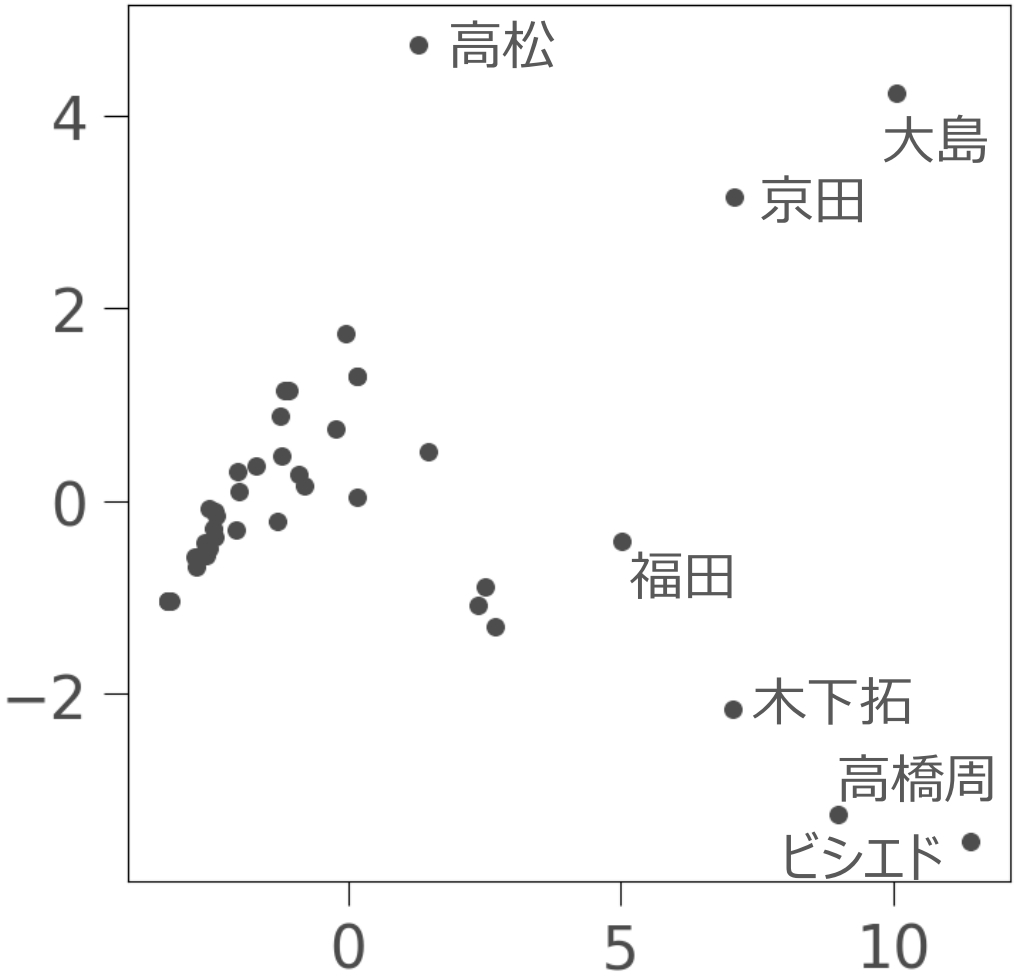
元のデータ（3次元）

次元圧縮後（2次元）

選手の特徴（主成分分析の結果）

打撃データの主成分分析

打撃22項目の主成分分析



各選手を、打撃22項目のデータに従い、22次元空間内の点として配置

↓

主成分分析で次元を圧縮

↓

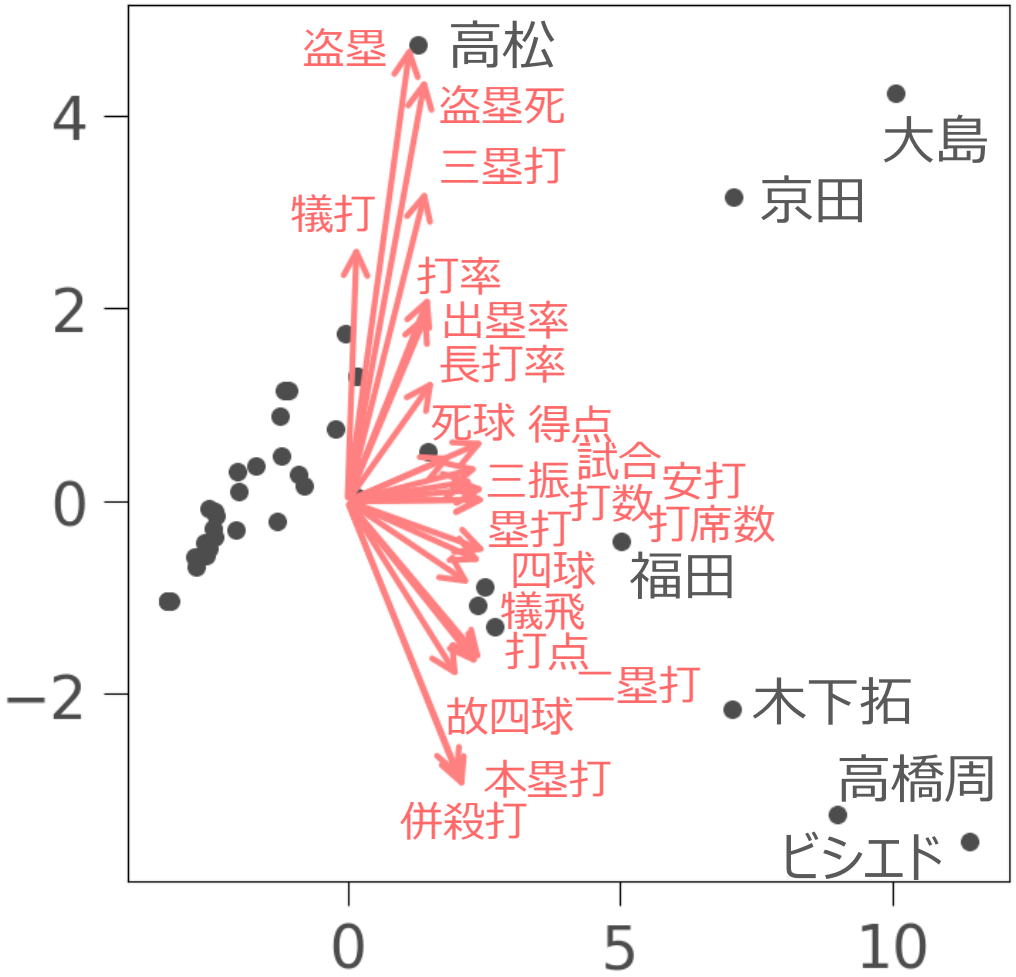
各選手を、2次元平面上の点として配置

- 点の位置が近い選手同士は、似たようなタイプ（成績）の選手。
- ただ、このままだと、点の位置がどのあたりだと、どのようなタイプの選手なのか、よく分からない。
- そこで、元の22次元空間の22本の軸も、空間と一緒に潰して、2次元平面上に図示（次ページ）。

選手の特徴（主成分分析の結果）

打撃データの主成分分析

打撃22項目の主成分分析



各選手を、打撃22項目のデータに従い、22次元空間内の点として配置

↓

主成分分析で次元を圧縮

↓

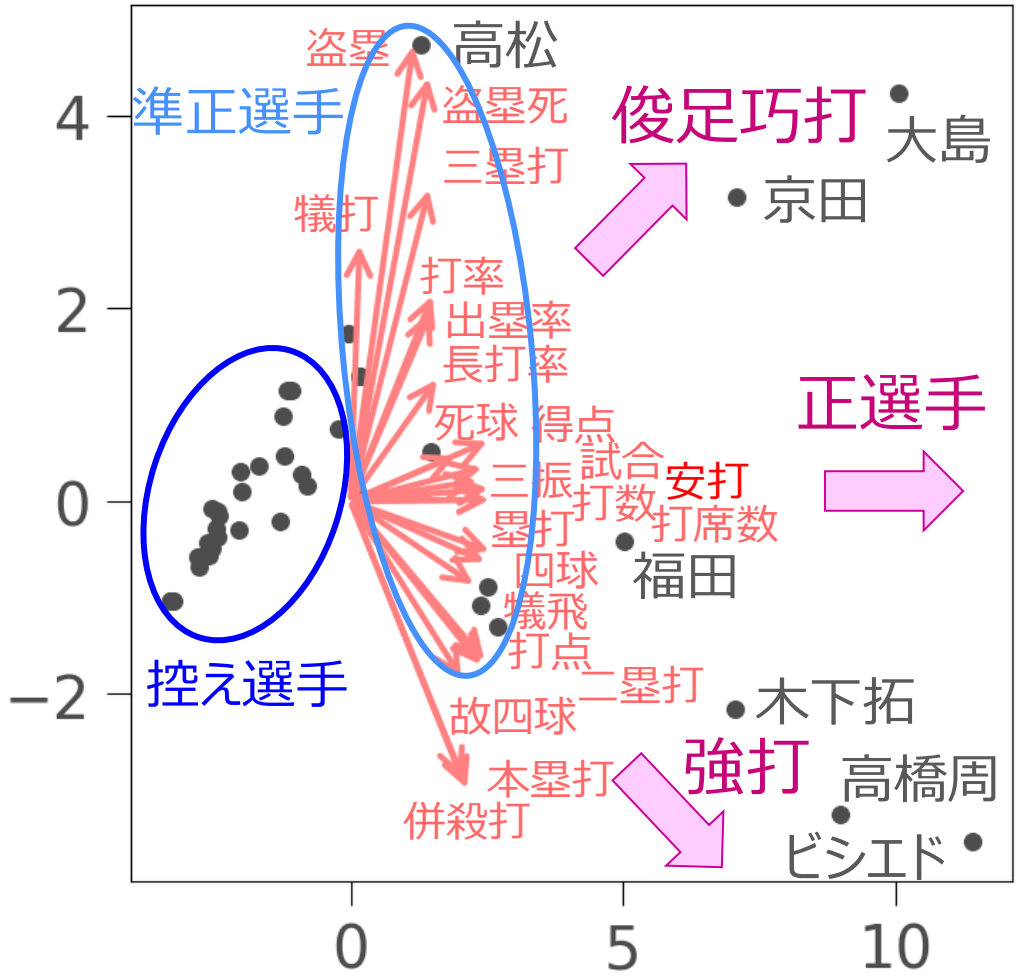
各選手を、2次元平面上の点として配置

- 図中の22本の矢印は、打撃22項目に対応。
- 原点（矢印の根本）から見て、各矢印の矢先の方向に向かうほど、該当する打撃項目の値が大きいことを意味する。

選手の特徴（主成分分析の結果）

主成分分析の結果の解釈

打撃22項目の主成分分析



試合数、打席数、打数に対応した矢印：右方向

↓

図の右側に位置する選手は、出場機会が多く正選手

打率、三塁打、盗塁に対応した矢印：右上方向

↓

図の右上に位置する選手は、俊足巧打の選手

本塁打、打点、犠飛に対応した矢印：右下方向

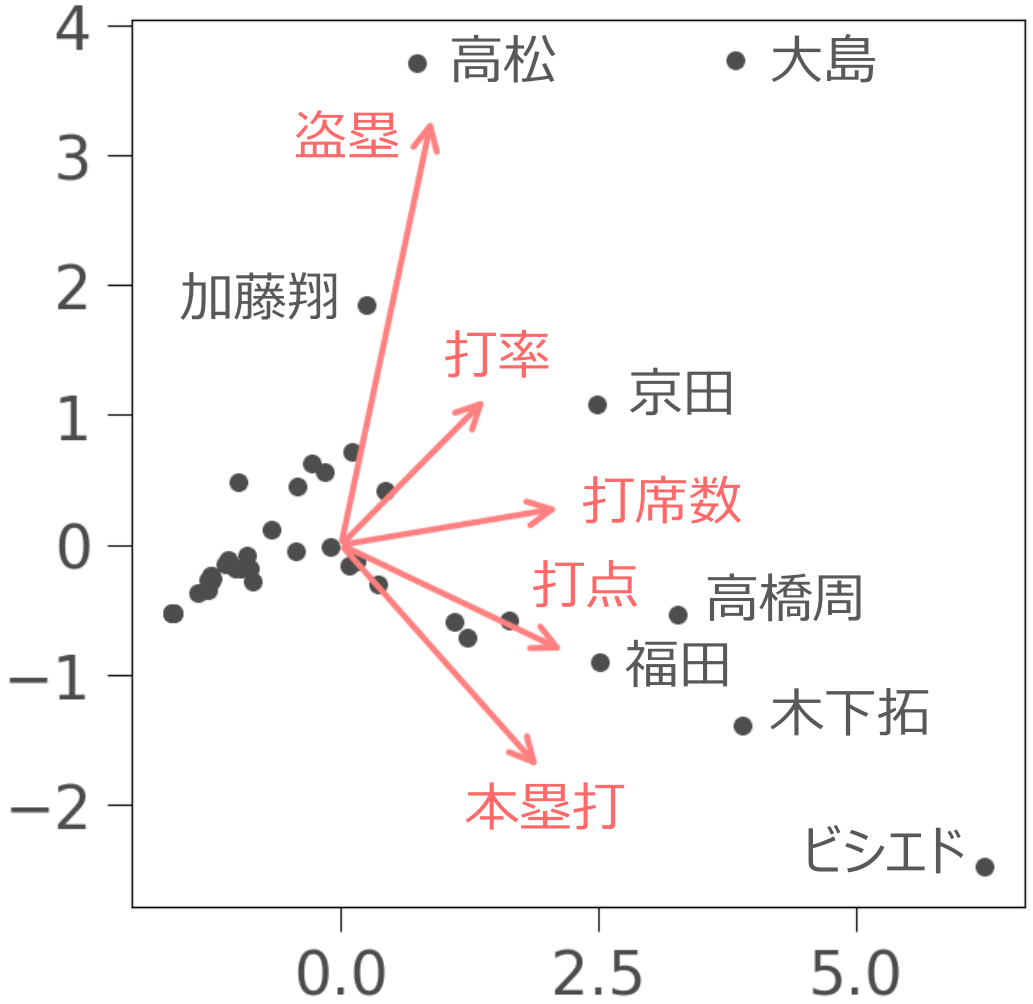
↓

図の右下に位置する選手は、強打が売りの選手

選手の特徴（主成分分析の結果）

打撃主要5項目を用いた主成分分析

打撃5項目の主成分分析



打者の特徴を分析する際、全22項目だと、似たような情報が重複するので、主要5項目（打席数、打率、本塁打、打点、盗塁）に絞り解析。

各選手を、打撃5項目のデータに従い、
5次元空間内の点として配置

↓

主成分分析で次元を圧縮

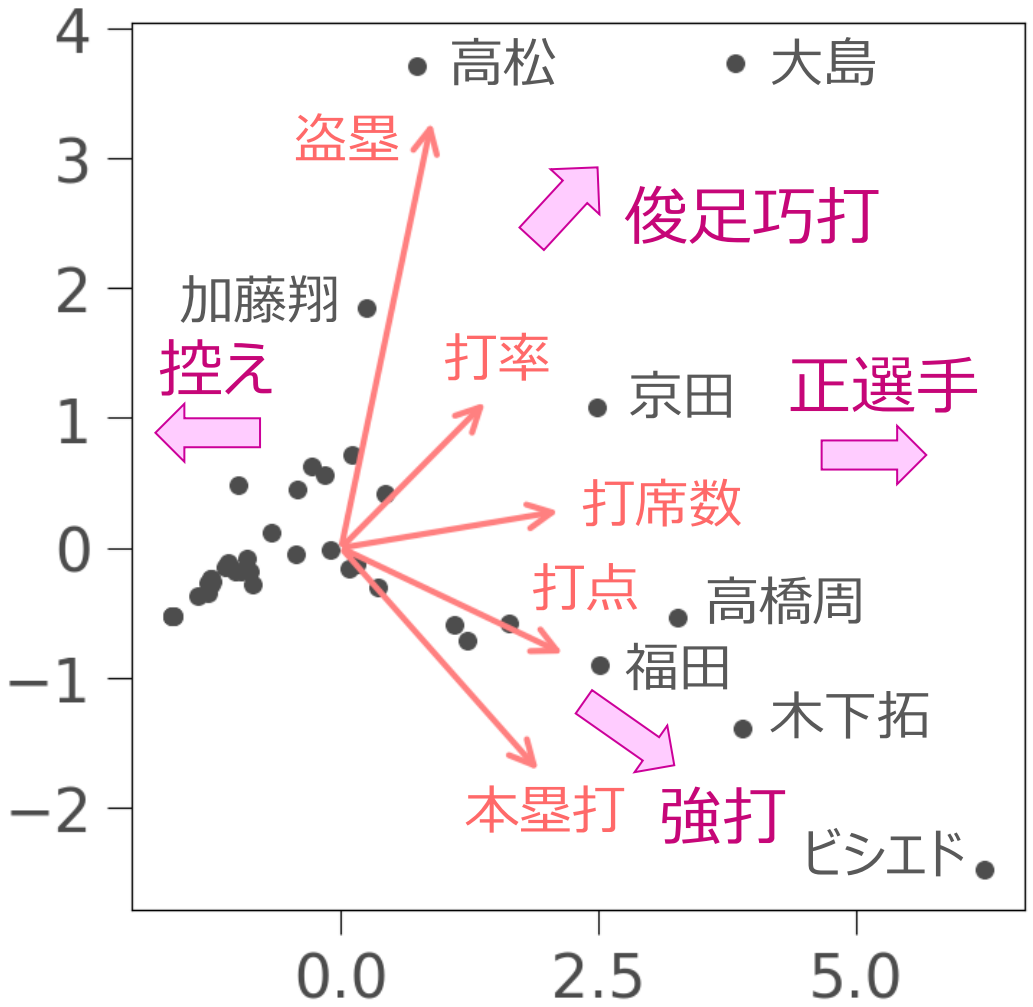
↓

各選手を、2次元平面上の点として配置

選手の特徴（主成分分析の結果）

主要5項目を用いた分析の解釈　22項目を用いた場合と、ほぼ同様の結果

打撃5項目の主成分分析



打席数に対応した矢印：右方向

↓

図の右側に位置する選手は、出場機会が多く正選手

打率、盗塁に対応した矢印：右上方向

↓

図の右上に位置する選手は、俊足巧打の選手

本塁打、打点に対応した矢印：右下方向

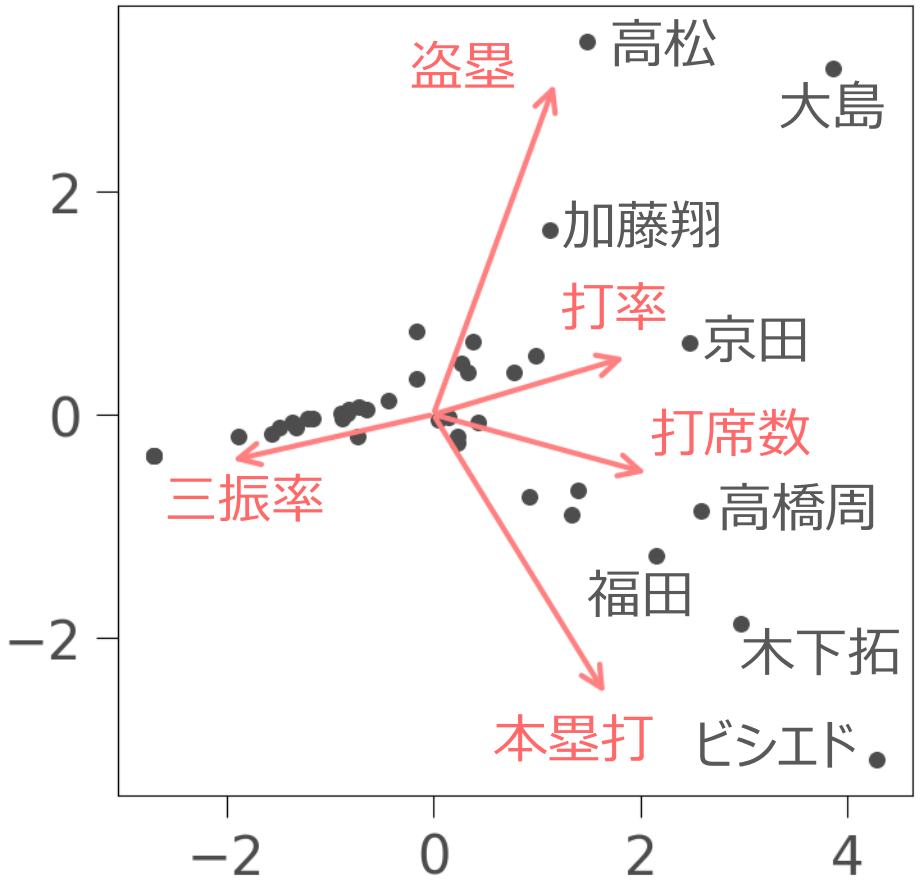
↓

図の右下に位置する選手は、強打が売りの選手

選手の特徴（主成分分析の結果）

三振のデータも用いて主成分分析

三振率を含む打撃5項目の分析



使用する項目を変えてみると？

試しに、打点をなくし、三振率（三振数÷打数）を使用（打席数、打率、本塁打、盗塁、三振率の5項目を使用）

⇩

三振率は、打率とほぼ逆の傾向にある（三振しやすいと、打率は低い）ので、矢印も逆向き

⇩

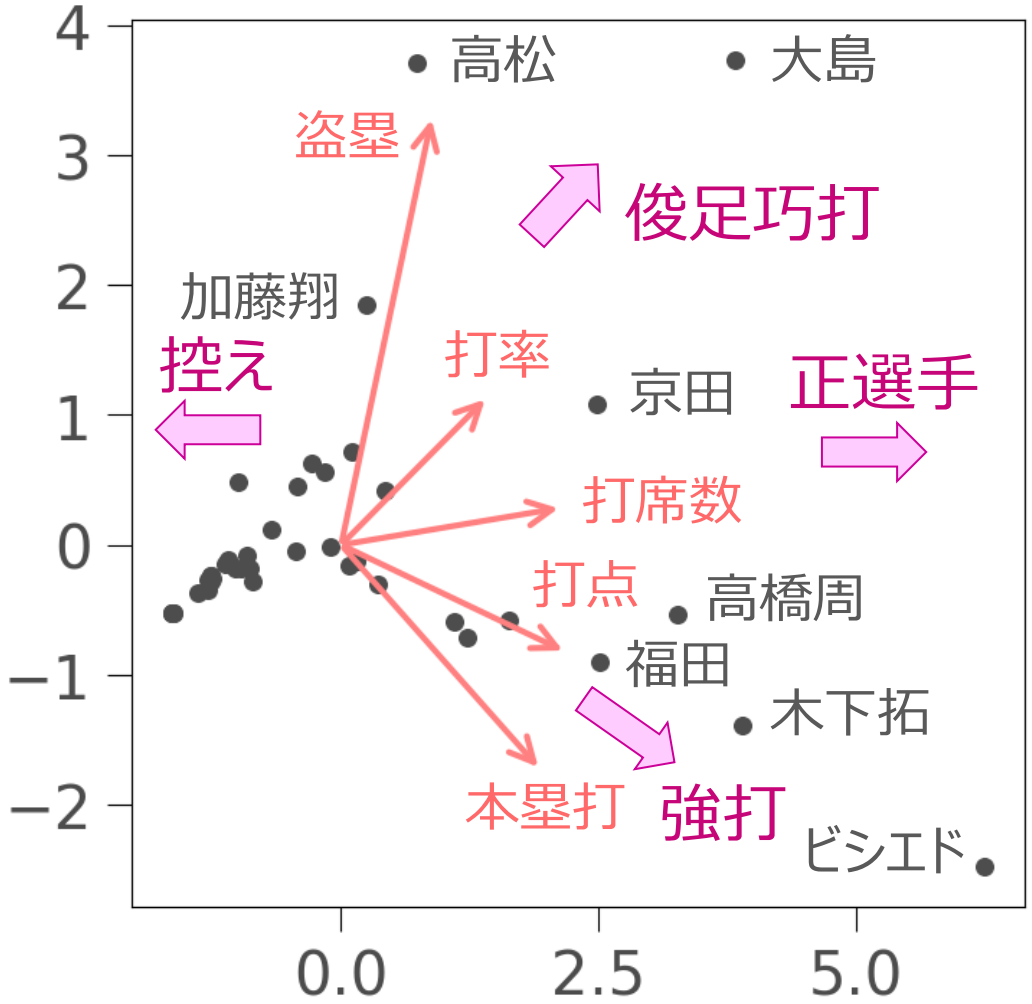
結果として、三振率の情報は、打率と重複

選手の特徴（まとめ）

選手の特徴の分析（結論）

2021年、ドラゴンズ所属選手のデータ
（一軍の試合で1打数以上記録した全43選手分）

打撃5項目の主成分分析



・打撃成績を元に、各選手を平面上に図示したら、出場機会（正選手 vs. 控え選手）と長所（巧打者 vs. 強打者）が反映された配置となった。

・打撃22項目のデータを全て用いても、主要5項目（打席数、打率、本塁打、打点、盗塁）に絞っても、ほぼ同様の結果が得られた。

データ出典：中日ドラゴンズ公式HP
シーズン打撃成績（2022年2月12日閲覧）
※現時点では、直近のシーズンのデータに更新済
https://dragons.jp/teamdata/batting.html

選手の特徴（おまけ）

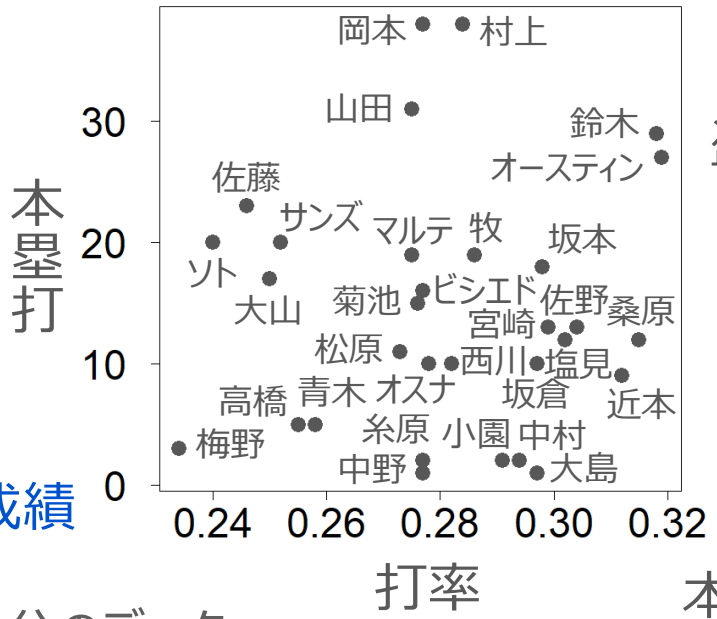
2021年プロ野球（セ･リーグ）の打撃成績の主成分分析

選手	打率	本塁打	盗塁	四死球	三振
オースティン	0.319	27	1	62	107
鈴木	0.318	29	9	87	71
桑原	0.315	12	9	35	64
近本	0.312	9	22	31	42
佐野	0.304	13	0	56	57
塩見	0.302	12	20	49	124
宮﨑	0.299	13	0	43	45
坂本	0.298	18	2	59	70
大島	0.297	1	16	43	51
⋮					

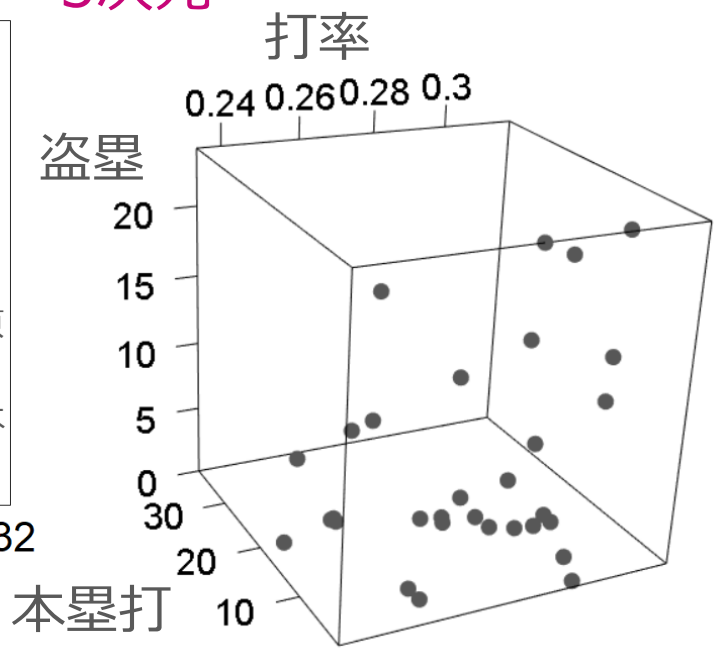
※2021年9月26日現在の成績
（最終成績ではない）

主力選手の特徴を比較するため、セ･リーグ6球団の打席数が多い選手を解析対象として、再度、主成分分析を実行

2次元



3次元



※打席数がチーム試合数の3.1倍以上の選手31名分のデータ

データ出典：日本野球機構（2021年9月26日閲覧） https://npb.jp/bis/2021/stats/bat_c.html

選手の特徴（おまけ）

2021年プロ野球（セ･リーグ）の打撃成績の主成分分析

選手	打率	本塁打	盗塁	四死球	三振
オースティン	0.319	27	1	62	107
鈴木	0.318	29	9	87	71
桑原	0.315	12	9	35	64
近本	0.312	9	22	31	42
佐野	0.304	13	0	56	57
塩見	0.302	12	20	49	124
宮﨑	0.299	13	0	43	45
坂本	0.298	18	2	59	70
大島	0.297	1	16	43	51

⋮

※2021年9月26日現在の成績　（最終成績ではない）

今回は各球団の主力選手（一定の打席数以上の選手）が解析対象のため、前述の控え選手も含めた解析とは違う項目を採用。例えば打席数は、対象となる選手間であまり差がないので、ここでは用いないことにした。

各選手を、打撃5項目のデータに従い、
5次元空間内の点として配置

↓

主成分分析で次元を圧縮

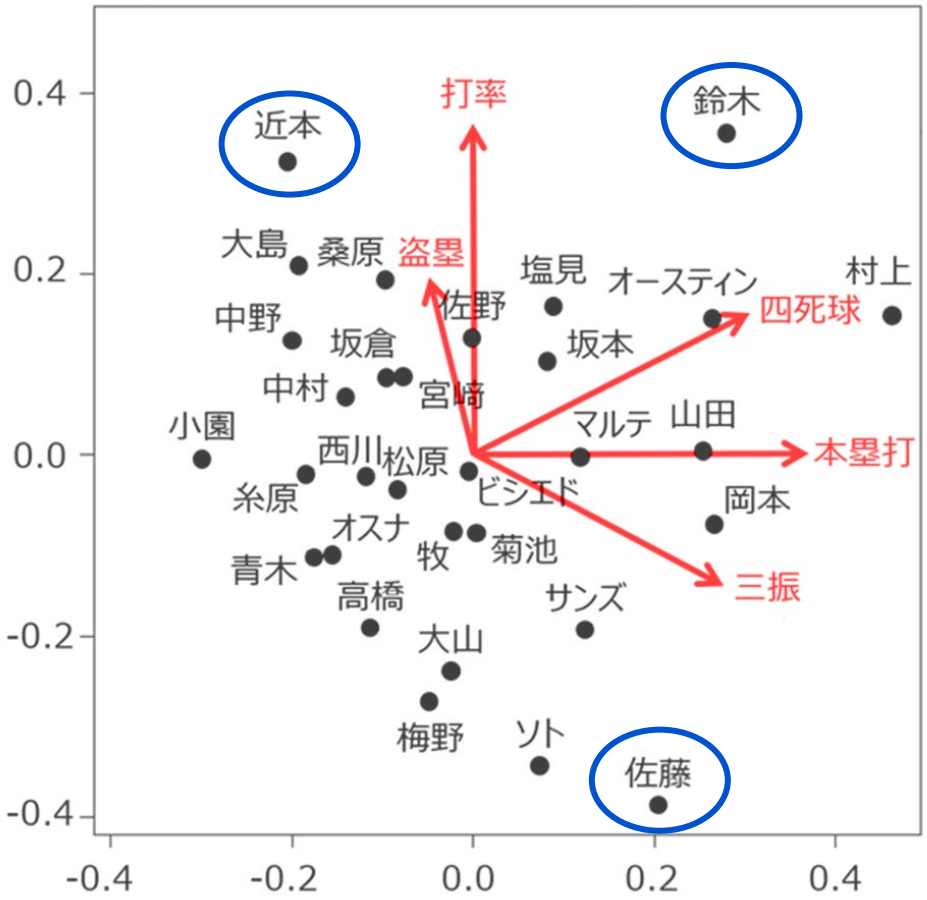
↓

各選手を、2次元平面上の点として配置

選手の特徴（おまけ）

2021年プロ野球（セ･リーグ）の打撃成績の主成分分析

5次元を2次元に圧縮



5本の軸（矢印）から考察（打者の一般論）

・打率と盗塁は似た傾向

・本塁打、四死球、三振は、比較的似た傾向

各点（黒丸）から考察（個々の打者の特徴）

・鈴木選手は打率が高く、本塁打も多い

・近本選手は打率が高く、盗塁も多い

・佐藤選手は本塁打が多いが、三振も多く、打率は低い

※大雑把な解釈であり、必ずしも厳密な議論ではない

データ出典：日本野球機構（2021年9月26日閲覧）
https://npb.jp/bis/2021/stats/bat_c.html

https://www.math.sci.ehime-u.ac.jp/~matsuura/

好不調？運不運？

統計的なバラツキ

～ 打撃成績がどの程度、運に左右されるか検証 ～

本塁打数（選手のタイプ）

ある打者（A選手）の本塁打数

A選手は、15年連続、レギュラーとして活躍。
各シーズンの年間及び月間の本塁打数は下記の通り。
この選手の活躍ぶりを、データから読み解いてみよう！

年	4月	5月	6月	7月	8月	9月	年間
1年目	1	2	4	4	2	4	17
2年目	4	1	3	3	7	4	22
3年目	3	6	4	2	5	5	25
4年目	2	8	2	1	1	4	18
5年目	2	5	3	5	2	3	20
6年目	3	2	2	4	2	5	18
7年目	4	6	6	5	3	2	26
8年目	3	5	1	1	1	3	14

年	4月	5月	6月	7月	8月	9月	年間
9年目	9	7	5	2	6	12	41
10年目	4	2	2	3	4	3	18
11年目	7	5	5	7	3	5	32
12年目	0	3	2	3	3	4	15
13年目	6	1	4	3	5	1	20
14年目	4	2	3	2	3	4	18
15年目	2	7	3	2	2	4	20
通算	54	62	49	47	49	63	324

本塁打数（選手のタイプ）

A選手の特徴

・1本も本塁打を打てなかった月もあれば、12本打った月もある　好不調の波が激しい

・3年目以降は、1年おきに、20本以上と19本以下の繰り返し　隔年で活躍

・通算の月間成績を見ると、6～8月の本塁打は少なめ　春秋に強く、夏に弱い

年	4月	5月	6月	7月	8月	9月	年間
1年目	1	2	4	4	2	4	17
2年目	4	1	3	3	7	4	22
3年目	3	6	4	2	5	5	25
4年目	2	8	2	1	1	4	18
5年目	2	5	3	5	2	3	20
6年目	3	2	2	4	2	5	18
7年目	4	6	6	5	3	2	26
8年目	3	5	1	1	1	3	14

年	4月	5月	6月	7月	8月	9月	年間
9年目	9	7	5	2	6	12	41
10年目	4	2	2	3	4	3	18
11年目	7	5	5	7	3	5	32
12年目	0	3	2	3	3	4	15
13年目	6	1	4	3	5	1	20
14年目	4	2	3	2	3	4	18
15年目	2	7	3	2	2	4	20
通算	54	62	49	47	49	63	324

本塁打数（選手のタイプ）

別の打者（B選手）の本塁打数

B選手も、15年連続、レギュラーとして活躍。
各シーズンの年間及び月間の本塁打数は下記の通り。
この選手の活躍ぶりを、データから読み解いてみよう！

年	4月	5月	6月	7月	8月	9月	年間
1年目	5	5	5	3	2	2	22
2年目	3	3	3	6	4	3	22
3年目	8	6	4	1	5	2	26
4年目	8	1	6	5	1	6	27
5年目	4	2	4	6	4	3	23
6年目	3	6	4	5	2	9	29
7年目	5	3	5	1	3	5	22
8年目	3	5	2	5	5	9	29

年	4月	5月	6月	7月	8月	9月	年間
9年目	3	3	4	2	5	9	26
10年目	3	3	5	3	7	5	26
11年目	6	4	7	3	8	3	31
12年目	2	3	4	4	3	4	20
13年目	3	8	4	4	1	2	22
14年目	2	4	6	3	5	4	24
15年目	6	7	4	4	4	1	26
通算	64	63	67	55	59	67	375

本塁打数（選手のタイプ）

B選手の特徴

・本塁打を2ケタ打った月はないが、1本も打てなかった月もない　好不調の波が小さい

・年間、最高で31本しか打っていないが、15年連続20本塁打以上　毎年活躍

・通算の月間成績を見ると7月が少なめ。それでも55本　季節間の調子の差が小さい

年	4月	5月	6月	7月	8月	9月	年間
1年目	5	5	5	3	2	2	22
2年目	3	3	3	6	4	3	22
3年目	8	6	4	1	5	2	26
4年目	8	1	6	5	1	6	27
5年目	4	2	4	6	4	3	23
6年目	3	6	4	5	2	9	29
7年目	5	3	5	1	3	5	22
8年目	3	5	2	5	5	9	29

年	4月	5月	6月	7月	8月	9月	年間
9年目	3	3	4	2	5	9	26
10年目	3	3	5	3	7	5	26
11年目	6	4	7	3	8	3	31
12年目	2	3	4	4	3	4	20
13年目	3	8	4	4	1	2	22
14年目	2	4	6	3	5	4	24
15年目	6	7	4	4	4	1	26
通算	64	63	67	55	59	67	375

本塁打数（統計的なバラツキ）

A選手とB選手の比較

A選手:お調子者（不調の時は当てにならないが、絶好調の時は頼もしい存在）

・1本も本塁打を打てなかった月もあれば、12本打った月もある。（好不調の波が激しい）
・3年目以降は、1年おきに、20本以上と19本以下の繰り返し。（隔年で活躍）
・通算の月間成績を見ると、6～7月の本塁打は少なめ。（春秋に強く、夏に弱い）

B選手:安定で堅実（爆発力はないが、いつでも信頼できる）

・本塁打を2ケタ打った月はないが、1本も打てなかった月もない。（好不調の波が小さい）
・年間、最高で31本しか打っていないが、15年連続で20本塁打以上。（毎年活躍）
・通算の月間成績を見ると7月が少なめ。それでも55本。（季節間の調子の差が小さい）

でも、実は両者の実力や特徴に「差」はなく、成績の差は、単なる偶然！

本塁打数（シミュレーション）

本塁打数のシミュレーション

A選手、B選手のデータとも、実在ではなく、下記の条件で人工的にデータを生成

・15年の間、毎年、4月から9月まで、各月ちょうど80回、年間480回、打席に立つ。
・打席に立った時、本塁打を打つ確率は、常に0.05。（好不調の波は一切ない）
・上記の同一の条件のもと、100選手分をシミュレーション。（設定に個人差は一切ない）
・A選手、B選手は、この100選手のうちの2選手分。

この設定だと、月間本塁打数の平均は、80×0.05＝4本。
しかし、確率現象なので、毎月ぴったり4本ずつ本塁打を打つわけではない。
本塁打が0本の月や、12本の月もあり得る。
サイコロを10回投げたのに、たまたま1の目が1回も出ないこともあるのと同じ。

偶然の結果なので、「好不調」ではなく「運不運」！

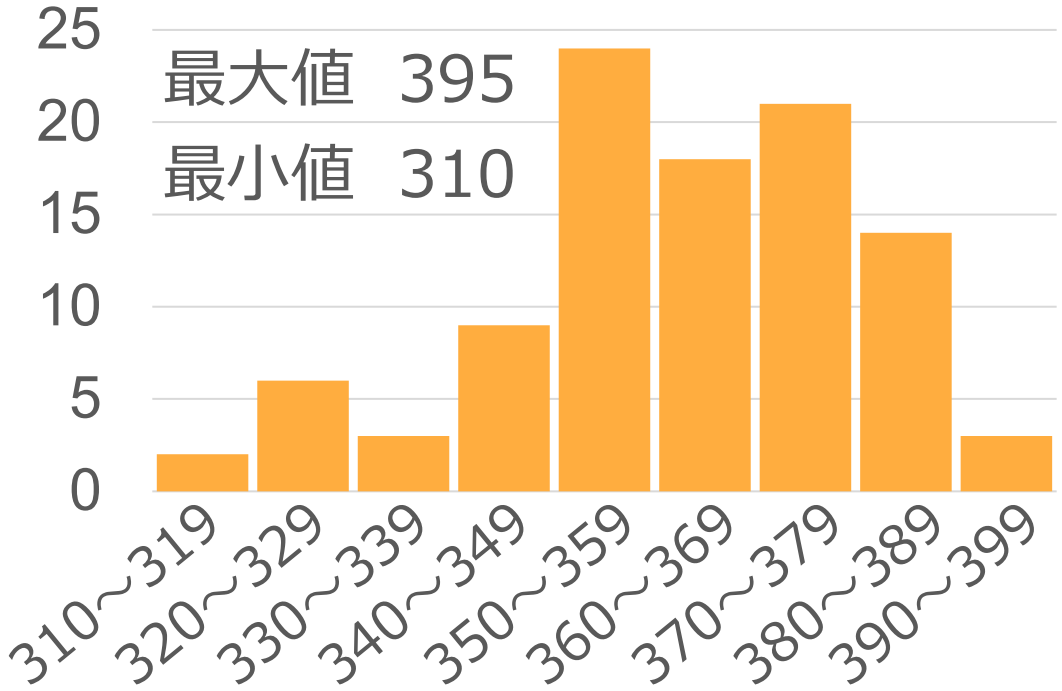
本塁打数（シミュレーション）

本塁打数のシミュレーション

100選手分のシミュレーション結果の集計（ヒストグラム）

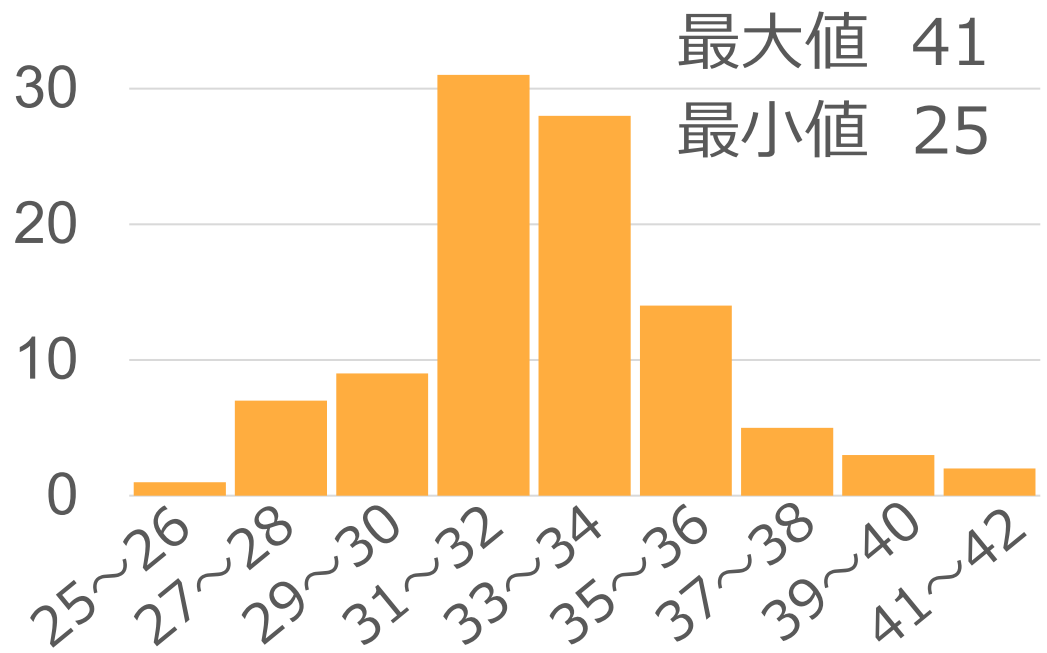
15年間の通算本塁打数の分布

平均で360本になるように設定



15年間における年間本塁打数の最大値の分布

年間本塁打の平均は24本になる設定



本塁打数（理論値）

理論値　　強打者（15年間の現役中、平均で通算360本塁打）を想定

・15年の間、毎年、4月から9月まで、各月ちょうど80回、年間480回、打席に立つ
・打席に立った時、本塁打を打つ確率は、常に0.05（好不調の波は一切ない）
・上記の同一の条件のもと、100選手分をシミュレーション（設定に個人差は一切ない）

この仮定のもと

・現役中（90か月の間）に、月間本塁打0本を経験する確率は約0.78
（不調でなくても、A選手のように、本塁打を打てない月があるのは普通）

・現役中に、月間本塁打12本以上を記録する選手が、100選手中、3選手以上いる確率は約0.91（A選手のように固め打ちをする選手がいるのは普通）

・現役中を通じ、月間本塁打が一貫して1～9本の範囲に収まる確率は、約0.12
（B選手のように月間本塁打数のバラツキが小さい選手は少数派）

本塁打数（まとめ）

統計的なバラツキ（まとめ）

設定

・毎打席、本塁打を打つ確率は0.05
・月80、年480、生涯7200打席
・好不調の波はなし
・個人差はなし

A選手（通算324本塁打）
月間本塁打0本もあれば12本もあり

B選手（通算375本塁打）
月間本塁打は常に1～9本

- 本塁打を打てるかどうかは確率現象のため、実力や調子が全く同じであっても、「運」に左右され、結果として、月間本塁打数には、バラツキが生じる。
- このバラツキは意外と大きいので、結果（本塁打の記録）だけ見て、好不調の判断や、打者のタイプ（気分屋 vs. 堅実者）の判別を行うのは、適切ではない（結果に一喜一憂しない！）。

全体のまとめ

この資料で学んだこと

- いろいろなデータ：数値データの他にも、テキストデータや画像データなどがある
- 疑似相関：2種類のデータの間に、相関関係（片方の値が大きいほど、もう片方の値も大きいという関係）があることと、因果関係（片方の値が、もう片方の値に影響を及ぼすという関係）があることとは、別物
- 線形回帰分析：2種類のデータの散布図（片方の値を横軸に取り、もう片方の値を縦軸に取ったグラフ）を、直線で近似（3種類以上のデータにも拡張可能）
- 非線形回帰分析：2種類のデータの散布図を曲線で近似
- 主成分分析：多くの項目からなるデータの特徴を要約し、平面上に可視化
- 統計的なバラツキ：確率現象においては、条件が一定でも、結果にバラツキが生じる。そうした、単なる「偶然の産物」に対して、特別な「意味」を見いださないように留意が必要