



データマイニング実践ガイド

第1回

目次

1. 「統計学的に」とは？ぴーち（p値）の誘惑
2. 実はとってもシンプル、機械学習の2大スター
3. モデル解析という本質的な考え方
4. エベレスト登頂の苦しみ、分析の裏側
5. 統計学の意外な得意技
6. 実例：クラスタリング
7. 実例：スコアリング
8. なぜデータ分析は役に立たないのか？
9. 本当に大切なたった一つのこと

イントロダクション

今日、皆さんが期待していると思っていることのイメージ ↓



長野オリンピック
スキージャンプ
日の丸飛行隊
金メダルをとった様子

<https://matome.naver.jp/odai/2139256005601394801/2139256886607826403>
より一部抜粋。

イントロダクション

実際に今日お話しすることのイメージ ↓



地獄絵図
剣山を登らされる様子

<https://blogs.yahoo.co.jp/fumin23wts/folder/562129.html?m=lc&p=1>
より一部抜粋。

イントロダクション

今日、この時間で学べること

1. 成功に近道はない、汗をかいた分しかお金はもらえない
2. 真剣に考えることをしなければ解決策などない
3. **本当にデータ活用をするということの覚悟と大切さ**

一見データマイニングとは関係なさそうだが???

1. 「統計学的に」とは？ぴーち（p値）の誘惑

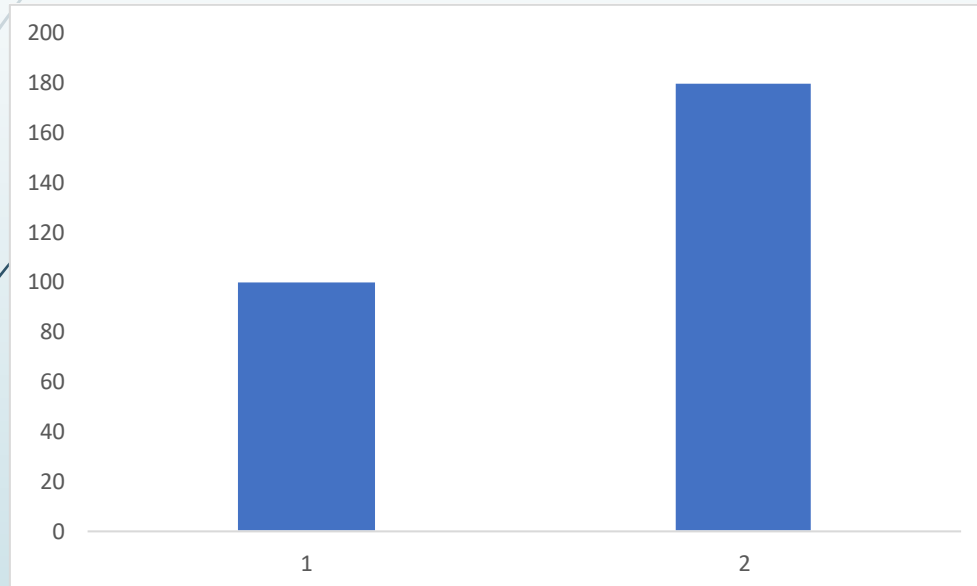
まず「はやりの言葉」に惑わされないように、以下の用語は一旦すべて忘れ去りましょう。以下の用語を区別することは重要ですが、それを行うことであなたの行動はどう変わりますか？

- ビッグデータ
- AI（人工知能）
- マシンラーニング（機械学習）
- ディープラーニング（深層学習）
- データマイニング
- 多変量解析
- 統計学、確率論 etc.

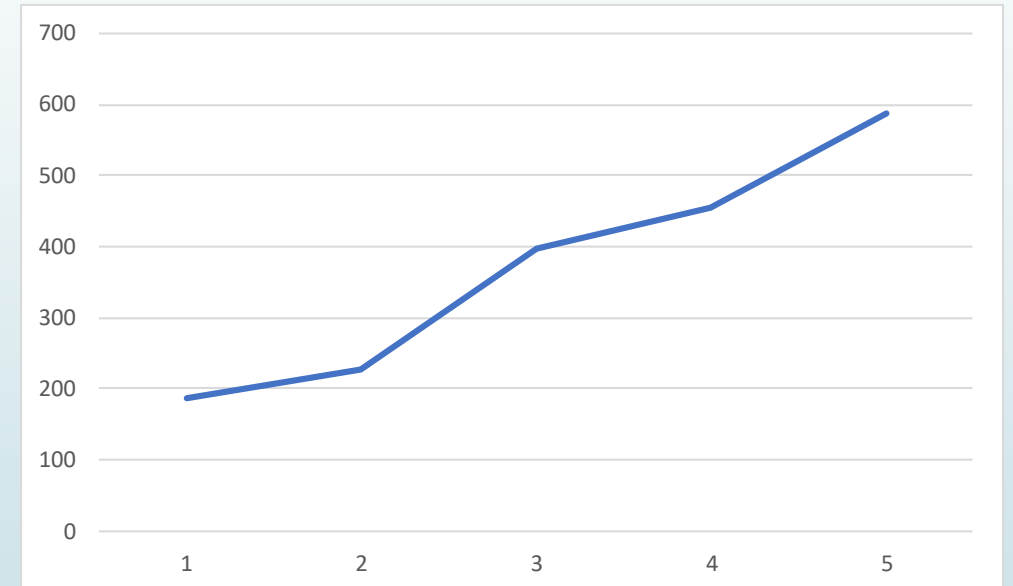
→もっと大切なことを理解しましょう

1. 「統計学的に」とは？ぴーち（p値）の誘惑

おそらく皆さんが想像する「統計学的に」というイメージ：



→ * * の方がいい



→右肩上がりに増加

1. 「統計学的に」とは？ぴーち（p値）の誘惑

あるいはこういうの：



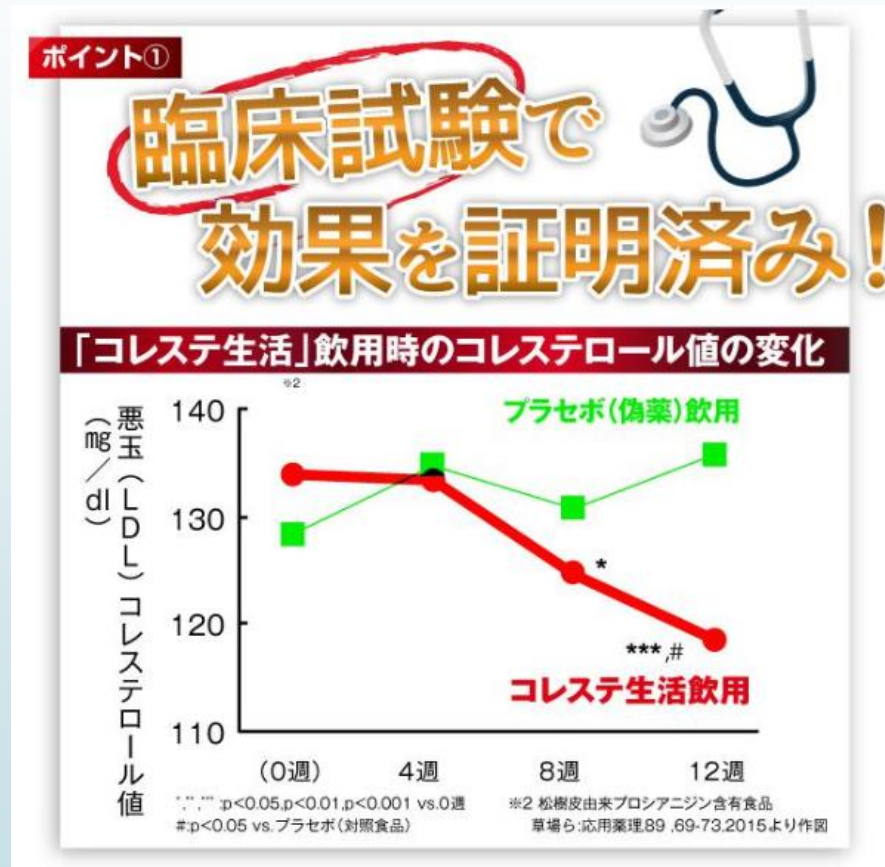
→サイの目が**な確率



→**回連続で勝つ確率

1. 「統計学的に」とは？ぴーち（p値）の誘惑

もしかしたら以下のような広告などを見かけたことがあるかもしれません：



<http://sharara-life.org/dmj-koresute-lifesappri>
より引用しました。

説明の例示のため、商品をすすめたり、効果の主張を促す意図はありません。

1. 「統計学的に」とは？ぴーち（p値）の誘惑

よく見ると色々なグラフの端っこに

*** $p < 0.05$

のような表現が書かれていたりします。

ちょっと詳しい人は、

- ・ pの値が小さいほど効果がある
- ・ アスタリスク (*) の数が多いほど効果が高い
- ・ 統計学的に差があることを示す記号

などということを知っているかも知れません。

1. 「統計学的に」とは？ピーチ（p値）の誘惑

私たちは意外と「うかつに」あまりよく知りもしない言葉を、何となく使っているかもしれません。今日は改めて、何となく使っている言葉をきちんと理解してみましょう。

【よく使ってしまう言葉】 ※どういう意味なんでしょう???

- ・ 統計学的に差がある（効果がある）
- ・ 統計学的にほとんど起こりえないことである
- ・ 統計学的に正しい答えが得られない可能性がある etc.

1. 「統計学的に」とは？p値（p値）の誘惑

そもそも統計学とはどういう学問なのでしょう。ここで紹介する「概念」を理解することで、統計学的なものを見方ができるようになります。一見、関係なさそうな話題にみえるかもしれませんが、とても大切なことです。

■重要なキーワード ※今日は難しい説明はしません。

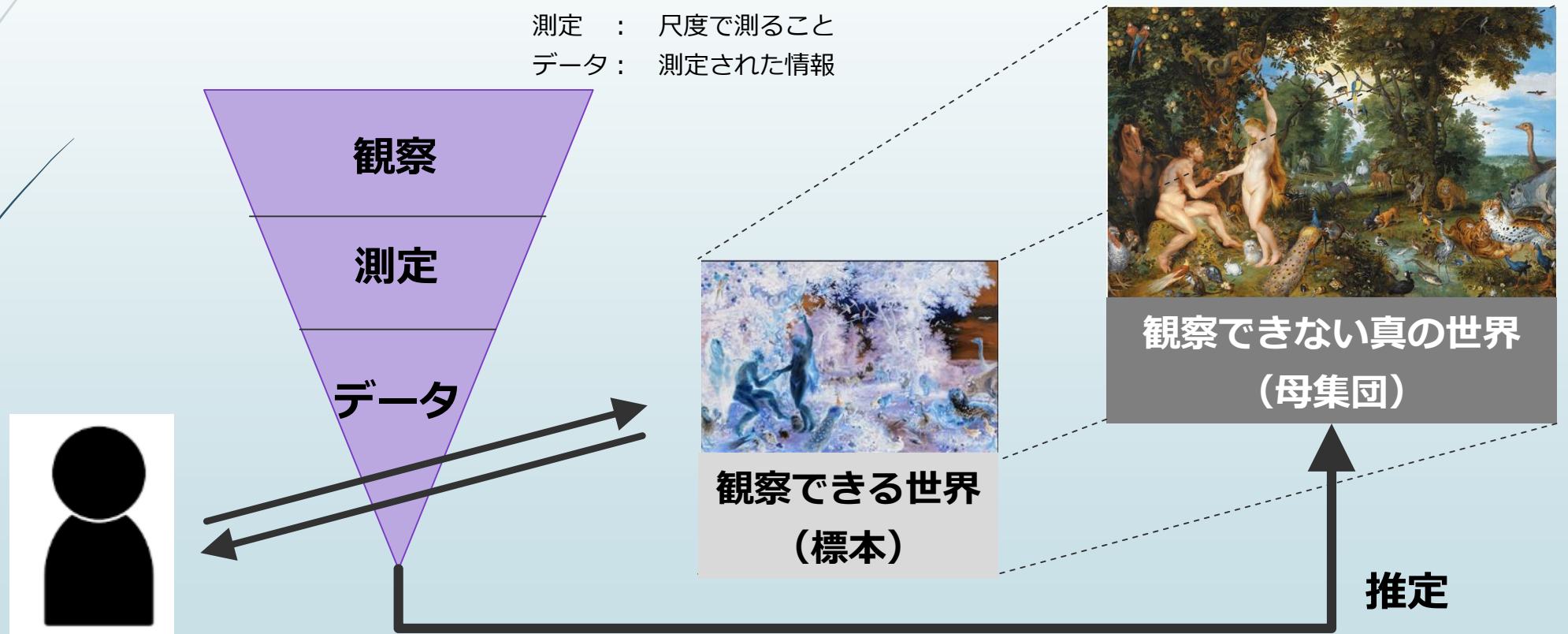
- ・母集団と標本
- ・観察、測定、データ
- ・帰納（確からしさの増大）
- ・推定

1. 「統計学的に」とは？ピーチ（p値）の誘惑

母集団と標本；観察、測定、データ；帰納（確からしさの増大）；推定を一枚の絵にして表現してみましょう。

https://www.lets-bible.com/fall/myth_b.php
より画像を引用。

観察：目で見ること
測定：尺度で測ること
データ：測定された情報



1. 「統計学的に」とは？ピーち（p値）の誘惑

我々が日常的にやっている例をいくつかあげてみましょう。

- ・ あー、この飲み屋の客層はオッサンばかりだな。
- ・ 日本の飲食店はどこもキレイだよね。
- ・ わたし、このブランドの柄がすごい好きなの。
- ・ 100円ショップはお得。
- ・ 今年の冬はいつもより雪が多いな。 Etc.

1. 「統計学的に」とは？ぴーち（p値）の誘惑

ひとつだけ、具体的に考えてみましょう。

- ・あー、この飲み屋の客層はオッサンばかりだな。

観察 : ある飲み屋にいた客を見た

測定 : (頭の中で) 見た目が40~50代に見える人の数を数えた

データ : (頭の中で) すべてのお客の数に対しての割合を計算した

※データを解析して推定した。

真実かどうか分からない。

→なぜ「真実かどうか分からない」のでしょうか？

ちょっと考えてみてください。

1. 「統計学的に」とは？ピーチ（p値）の誘惑

真実かどうか分からない理由はいくつか考えられそうです。

- ・ たまたま、その日だけオッサンが多かったのかもしれない。
- ・ オッサンと判断した人数が数え間違っていたのかもしれない。
- ・ その人がオッサンと勝手に思っただけかもしれない。

1. 「統計学的に」とは？ピーチ（p値）の誘惑

これを換言すると以下のように表現できます。

- ・ たまたま、その日だけオッサンが多かったのかもしれない。
→ **正しく観察、測定されたが、場所と日付が適切でなかったかもしれない。**
（母集団の流動性）
- ・ オッサンと判断した人数が数え間違っていたのかもしれない。
→ **測定が適切でなかったかもしれない。**
（正しい尺度、測定方法。測定誤差。）
- ・ その人がオッサンと勝手に思っただけかもしれない。
→ **対象の定義が妥当なものでなかったかもしれない。**
（構成概念妥当性）

1. 「統計学的に」とは？ピーチ（p値）の誘惑

難しい用語がでてきてしまったので、わかりやすく言うところになります。

- ・たまたま、その日だけオッサンが多かったのかもしれない。
→正しく観察、測定されたが、場所と日付が適切でなかったかもしれない。
(母集団の流動性) →**1階と2階、日付によって違う。**
- ・オッサンと判断した人数が数え間違っていたのかもしれない。
→測定が適切でなかったかもしれない。
(正しい尺度、測定方法。測定誤差。) →**測り方によって違う。**
- ・その人がオッサンと勝手に思っていただけかもしれない。
→対象の定義が妥当なものでなかったかもしれない。
(構成概念妥当性) →**何を測ろうとしているのかが曖昧。**

1. 「統計学的に」とは？ぴーち（p値）の誘惑

我々がデータを得ようとしたとき、そこには「観察」と「測定」という行為が発生します。これには必ず「バラつき」が生じるということが想像できると思います。実は簡単に測定できているものでも、絶対に真の値を測定することはできません。

■ 統計学の根本的な思想

- **母集団の値を観察、測定することは絶対にできない**
(現実的にも不可なことが多い)
- **測定にはバラつきが生じる**
(測定方法が適切であったとしても)

1. 「統計学的に」とは？ピーチ（p値）の誘惑

「統計学的に」というと、とにかく、たくさんのデータを集めて、代表的な値が得られれば「確率的に」何か正しいと言えそうだと主張できることとされていることが多いかもしれません。

しかし、統計学の本質は

・バラつきを説明する

ということ部分にあります。

※これについては、後ほど説明することになります。

2. 実はとってもシンプル、機械学習の2大スター

さて、少し気分転換をするために「はやりの言葉」を少し網羅しておきましょう。改めて機械学習などというと、目新しそうな気がしますが、これらは多変量解析という枠組みの中で何十年も前から存在しているものです（この手の手法の源流は1900年くらいから盛んに研究され始めている歴史があります）。

- ・ニューラルネットワーク
- ・サポートベクターマシン (SVM)
- ・ディシジョンツリー (決定木)
- ・自己組織化マップ
- ・アソシエーションルール
- ・ロジスティック回帰分析
- ・重回帰分析
- ・k-means、X-means (クラスター分析)
- ・レコメンド
- ・主成分分析
- ・因子分析
- ・正準相関分析
- ・コレスポンデンス分析
- ・数量化理論
- ・多次元尺度構成法
- ・ベイジアンネットワーク
- ・マルコフ連鎖モンテカルロ法
- ・状態空間モデル

2. 実はとってもシンプル、機械学習の2大スター

色々な手法があると、私たちが想像もできないような特別なアウトプットが得られるような気がしてきます。しかしながら（残念ながら？）、**すべての手法は以下の2点に集約されます。実質的なアウトプットは大差ありません。**

1. ある現象を説明（予測）する

重回帰分析、など。 →アウトプット：係数表、予測スコア

2. 複雑なデータを要約（整理）する

クラスター分析、など。 →アウトプット：重心表、クラスター番号

2. 実はとってもシンプル、機械学習の2大スター

注意

手法のことを大げさにアピールする人は詐欺師である可能性が高いです。
(自分が詐欺師であることにさえ気づいていないかも・・・)

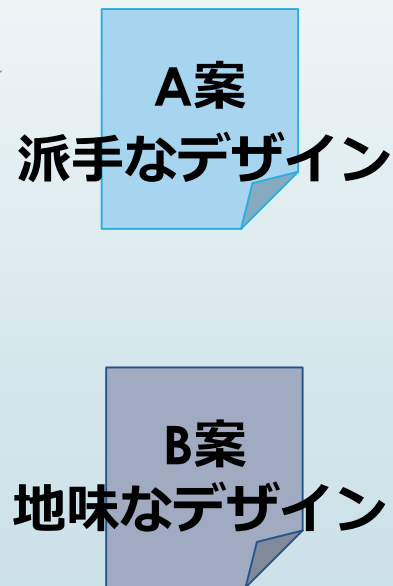
【典型的な恥ずかしい例】

- ・「でーぷらーにんぐ」使ってます
- ・「べいず」使ってます

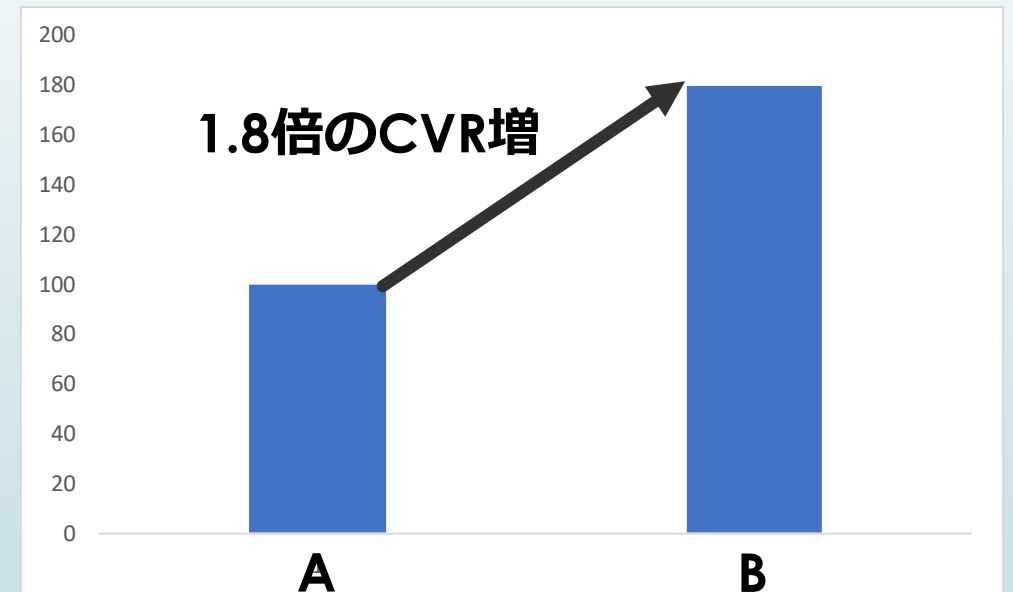
※素人が日本刀を持ったらケガをするだけです。逆に達人であればビニール傘でも凶器になります。上記のような人は「俺んちのとーちゃん政治家なんだぜ」と自慢する小学生のようなものです。作戦上、あえて、はやりの言葉を使っている可能性はありますが、、、、

3. モデル解析という本質的な考え方

日本刀を持って怪我をしないように（というか人を殺めないように）、本質的な考え方を学んでおきましょう。ここでは、webの世界でもよく使われるようになった「ABテスト」というものを例に挙げて説明してみます。



CVR = コンバージョンレート
通販サイトでいえば「購買率」のこと。



3. モデル解析という本質的な考え方

そもそも「差がある」とはどういうことでしょうか？

例えば、古典的な発想ですが「男性と女性の身長の平均値は違う」というのは直感的に納得しやすいことのような気がします。

統計学では「差がある」という表現は使いません（というか使ってはいけません）。「差が認められた」という表現を使いますが、ここに本質が隠されています。

3. モデル解析という本質的な考え方

先に結論をいってしまうと：

- ・ 差が認められるとは、（AとBとでCVRの差が認められるとは）
- ・ **Aグループ内、Bグループ内それぞれでの「平均まわりのバラつき」が小さく、**
- ・ **Aグループの平均とBグループの平均が大きくなっている状態をさす。**

※専門的には「郡内変動が最小かつ群間平均が最大である」という状態のことをいいますが、この状態を「検定統計量」という指標で表現します。帰無仮説が正しいときに、この検定統計量が得られる（正しく帰無仮説を棄却できる）確率のことをp値といいます。母集団は決して観察、測定できないので「たまたま観察、測定できた標本」から、確率的に真の値を推定することしかできません。そのため「差があるかどうかという真実を語ることはできないのです。

3. モデル解析という本質的な考え方

いきなり唐突すぎたので、モデル解析ということのイメージをお伝えしましょう。実は、私たちは日常で色々なモデル解析を無意識にやっています。その一例をあげてみます。



3. モデル解析という本質的な考え方

これは冬の代表的な星座（別に夏でも見えますが）、オリオン座です。実際には何の意味もない（はず）の星の位置を「**オリオン座**」という**モデルを当てはめる**ことで、**点の羅列が説明**されています。



3. モデル解析という本質的な考え方

もっと理屈っぽくいえば、2次元の平面上に布置されているように認知できる星の羅列から、星の明るさ、星の大きさ、隣接する位置の距離間隔、特定の星と星の相対位置を1つの集合体として認識することができる、ということです。



3. モデル解析という本質的な考え方

星の明るさ = {12, 3, 45, 120, ...} 例えばルクスという輝度の単位がある

星の大きさ = {5, 7, 3, 4, 9, ...} 円ととらえるなら円の面積とか

隣接する位置の距離 = {10.2, 5.8, 23.3, ...} ユークリッド距離というものがある

相対位置 = {2.5, 55.3, 109.9, ...} どこか1つの基準からの相対距離

※上記の数値はデタラメ



一見、我々が何となく認知しているものは「データで表現」することができる。このデータの集合が何かを「外的基準」として与えられることで、データに意味を持たせることができる。

3. モデル解析という本質的な考え方

要するに、

- ・データに対して
- ・何かを当てはめることによって説明がつく

ということは直感的に（無意識に）やっていることであり、統計学とは、

- ・ある統計モデルを当てはめることで、
- ・何かの差を検出できたり、予測や判別することができる
- ・ただし、データはつねに「バラつき」をもっている
- ・このバラつきを説明できたかどうか重要

とだけ覚えておきましょう。

3. モデル解析という本質的な考え方

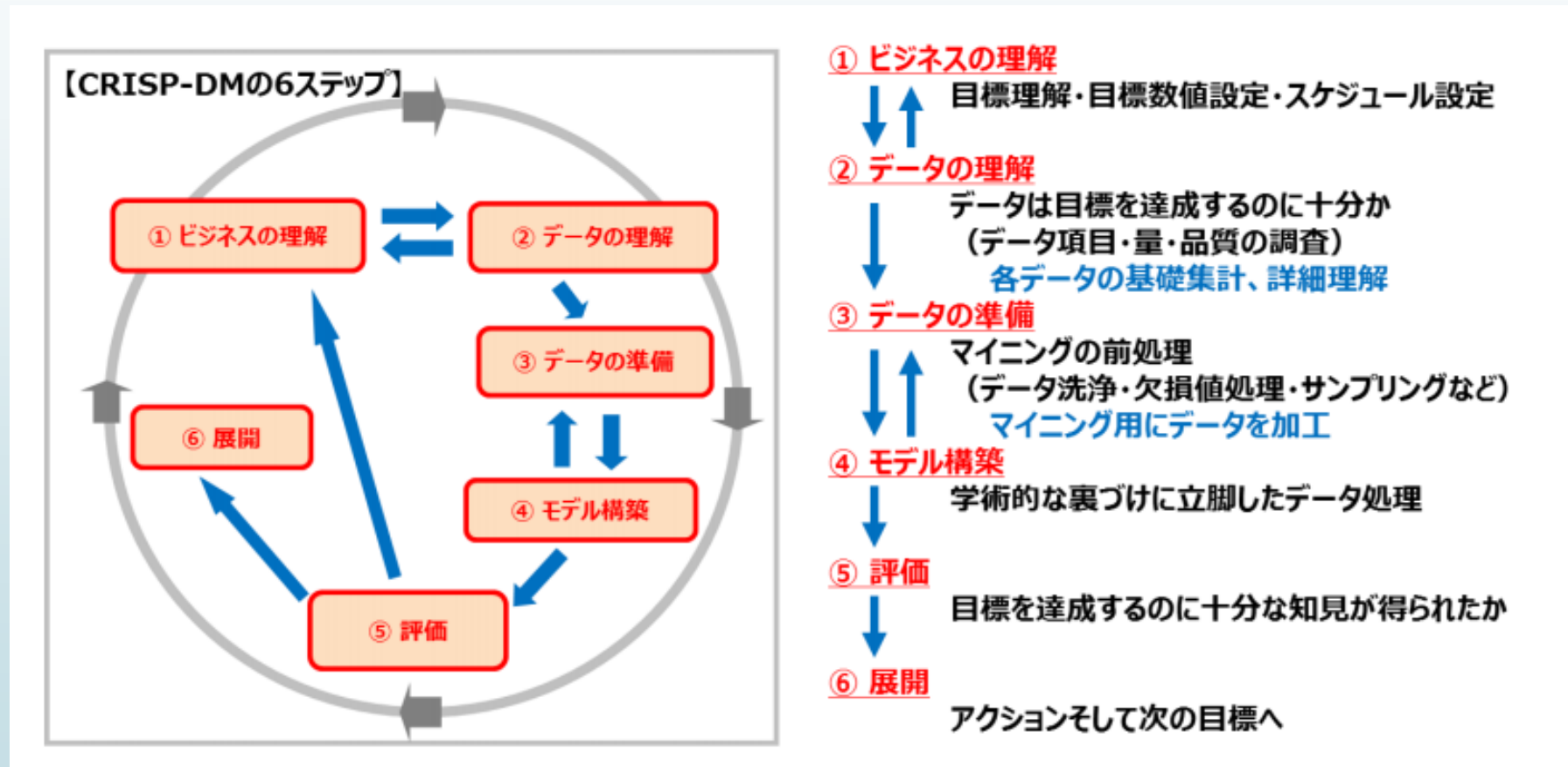
文章だけでは何を言っているのかイメージできないと思いますので、以下のページでその具体的な様子を見てみましょう。

シリウス先生の心理統計学： 分散分析の基本原理

http://daas.la.coocan.jp/GLM/hosoku_2_anova.htm

4. エベレスト登頂の苦しみ、分析の裏側

実は分析するということ（あるデータに対して分析手法を適用して結果を得ること）は、さほど難しいことではなかったりします。ビジネスシーンでデータ分析を行う場合、どのような流れで進めていけばよいかという系統立った方法論があります。



4. エベレスト登頂の苦しみ、分析の裏側

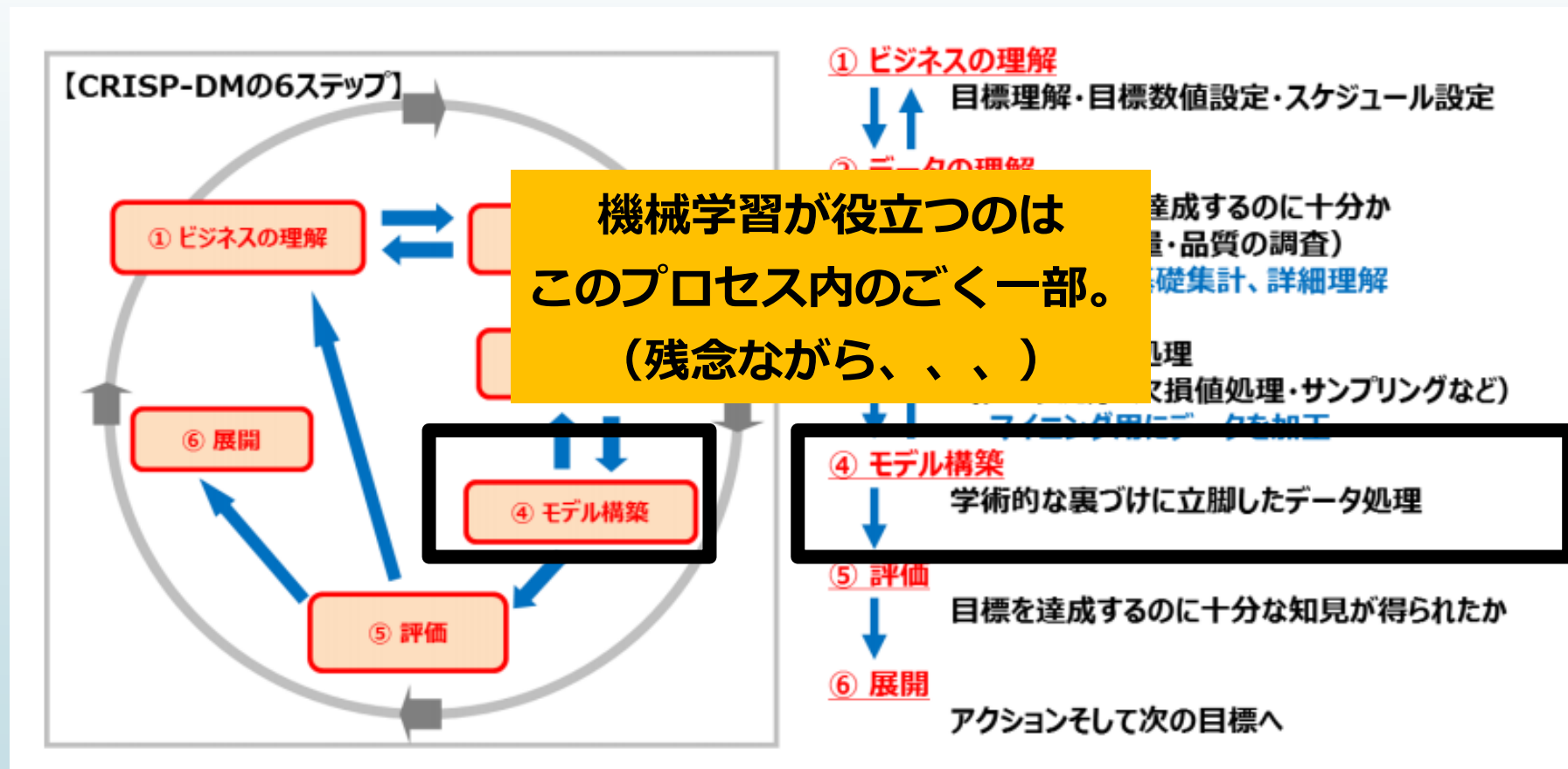
CRISP-DMは厳密な手順がPIMBOK（システム開発の手順書のようなもの）のようにまとめられています。IBMはSPSS Modelerの解説書という位置づけ（？）でCRISP-DMの手順を62ページにもわたる資料で説明しています。

IBM SPSS Modeler CRISP-DM ガイド

<ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/ja/CRISP-DM.pdf>

4. エベレスト登頂の苦しみ、分析の裏側

現在、少なくとも日本で（呼び方は何でもいいですが）機械学習やら何やらといわれている技術は、実際のデータマイニングプロジェクトにおいては**一過程の一部を手助けするものにすぎません。**



4. エベレスト登頂の苦しみ、分析の裏側

では実際にどれだけの工程をふまなければならないのかを紹介してみましよう。ここであげる工程は一般的なものにすぎず、実際にはさらに細かい工程が存在します。またそれは常に流動的に動かされるものです。

番号	ステップ	統合
1	ビジネスの理解	会社もしくは事業での戦略の把握 多くは有報から情報収集、推察する 課題の列挙および整理 とにかくアイデアを出す 会社もしくは事業の戦略上、有意義なものを選択 具体的な業務フローの把握、図示化 具体的なPJ着手ポイントを決める ビジネスインパクト×分析パフォーマンス PJ着手ポイントのゴールを明確にする ビジネス上達成したい目的・分析により得られる成果 データイメージの作成 分析アプローチ概要の作成 分析要件の一次設定(bodais 1~5番のやつ) 分析アウトプットの具体的図示 分析工程の作成、工数算出 分析スケジュールの作成

4. エベレスト登頂の苦しみ、分析の裏側

2	データの理解	使用可能なデータの把握(リスト確認) データの初期入手 データの構成概念の整理、データテーブルの分類 導き出そうとする結果を説明する現象の整理 データテーブルイメージの具体的図示 データ整理(受領ファイル、カラム、変数加工仕様の策定、smallデータ; テーブル判定=M/Tr、変数判定=連/カテ) 基礎統計量の算出(統計値、分布)と生値照合 QA表の作成 データ処理フローの具体的図示 変数加工仕様の確定
3	データの準備	分析要件の二次設定(データの状況から具体的な集計期間、目的変数の定義等を決定する) データの一次加工(特定期間、特定カラムの抽出; データサイズの縮小) トランザクション集計 派生変数の作成(概念的にデータクレンジング含む) テーブルマージ モデル構築に用いる変数の分布チェック
4	モデル構築	データのランダムサンプリング(大規模な場合、分析用データセット作成) 決定木によるデータ構造の可視化、複雑な交互作用のチェック 個別変数の効果量算出(ex.AIC、F値ランキング) モデル初期変数の確定 モデル作成(説明:スコアリング or 整理:クラスタリング) モデルに対する考察 モデル選択(試行錯誤で有力候補を2~3個見つけ出すイメージ) 交差検証 予測結果の内容的妥当性のチェック(ex.スコアリングの場合...生値照合、平均正解率など)

4. エベレスト登頂の苦しみ、分析の裏側

5	評価	モデルの解析的視点による評価(ROC等;信頼性が高く説明力の高いものを選ぶ) 業務フローにおける改善効果の検討 実際の運用方法との適合性検討 モデルの内容妥当性の評価(効果指標等、業務フロー上運用可能なものを選ぶ) 最適モデルの決定(有力候補の中から、最終的に用いるものを決定) 分析によって得られた知見のまとめ
6	展開	分析結果の具体的な使用法を図示化 実業務上での適用可能性を検討 施策実施に向けての準備項目の列挙 テストマーケティング計画/実検証の実験計画 データ処理フローの具体的図示(精緻化して更新) 分析要件、分析手順のまとめ(再現性確保) システム運用に向けての条件整理

※例えばどんなデータが手元にあって、そのデータをどう組み合わせられるか、どういう加工をして分析に使うかといったことをまとめなければなりません。これでさえ基本スキルが習得されていなければ実行できないのです。

シリウス先生の心理統計学： 実践データサイエンス「4. データ整理」

http://daas.la.coocan.jp/jisen_datasci/jds_04_dataseiri.htm

5. 統計学の意外な得意技

最近、もてはやされている機械学習の手法のひとつにニューラルネットワークというものがあります。これは判別分析と呼ばれるものの1つで、「Yes / No」を判別するという古典的なものです。これをさらに高精度化したものがディープラーニングなどと呼ばれていますが、計算技術とコンピュータ技術によって効率かつ高速化されただけで、アウトプット自体は（精度という面を除けば）同じものです。

Data → Yes / No
Model



ニューラルネット（以前） → 最近

5. 統計学の意外な得意技

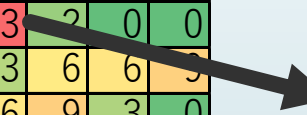
(余談)

なかなか想像しづらいかもしれませんが、画像というのは「0-1」の2値データが高次元で表現されたものです。何を言っているのか全く分からないでしょうからイメージを描いてみましょう。




0	0	12	23	2	0	0
0	1	2	3	6	6	5
10	4	8	6	9	3	0
2	6	7	5	6	2	6
3	9	9	0	9	2	9
5	6	9	8	6	9	9
4	3	10	6	4	8	3
8	9	4	3	2	6	1

画像とは「マス目」
に色の情報を番号で
表現したもの。



0	0	0	1	1	1	1
---	---	---	---	---	---	---



0	1	0	0	0	1	0
---	---	---	---	---	---	---

1つのマス目に入っている数値
は1行の0-1羅列で表現すること
ができる。

5. 統計学の意外な得意技

(余談)

あらゆる情報は「0-1」の「データ」によって表現することができます。このデータの羅列（人間にはこれが何なのかわからない）が「猫である」と機械が認識するのが「画像認識」という技術です。本日はこの話を深掘りしません。

0	0	0	0	1	0	0	1	0	0	0	0
1	0	1	0	1	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1	1	0	1	0
1	0	1	0	1	0	0	1	0	1	0	0
0	1	1	1	0	0	0	0	1	1	1	0
0	0	0	0	0	0	1	1	0	1	0	0
0	0	0	1	1	0	1	1	1	0	0	1
1	0	1	0	0	0	1	1	0	0	1	0
1	0	0	1	0	1	1	0	1	0	1	1
1	1	0	0	0	0	0	0	1	0	0	1
0	0	1	0	1	1	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	1	0	1	0	1
0	1	0	0	1	1	0	1	0	0	0	0
0	0	0	1	0	1	0	0	0	0	0	0
0	0	1	0	0	1	1	0	1	0	0	0
1	0	0	0	0	0	1	1	0	0	0	0
1	1	1	0	1	1	1	0	0	0	1	1

=



5. 統計学の意外な得意技

統計学というと、どうしても「確率」というイメージが強いかもしれませんが。しかし、統計学の手法は数学からではなく、生物学や心理学といった分野での問題を解決するために編み出されたものが多いのです。先に紹介したABテストも、元々はフィッシャーという人が農場実験をするというところから発展していった考え方です。

■ 意外な得意技の代表例

- ・ 本当においしいと思っているケーキはどれか？ 【一対比較法】

http://daas.la.cocacn.jp/toukei_hosoku/paired_comparison.htm

- ・ 右手で練習すれば左手でもできる 【共分散分析】

http://daas.la.cocacn.jp/toukei_hosoku/series_shinrikenkyu/transfer_of_training.htm

- ・ 本当にいいメイド喫茶はどこだ 【相対度可視化法】

<http://www.msi.co.jp/splus/events/student/2010pdf/10kata.pdf>

5. 統計学の意外な得意技

これらの例から学ぶべきことは、手法そのものというよりも

- ・ **統計手法とは常に科学的に妥当かつ信頼のある評価をしたい**

という根源的なモチベーションから生まれている、ということでしょう。

(脇道) 妥当性と信頼性も実はみんな知らずに使ってしまったかも、、、

http://daas.la.cocacn.jp/toukei_hosoku/shinrai_and_datousei_kentou.htm

6. 実例：クラスタリング

機械学習の2大スターのところで以下のように書きました（以下再掲）。

1. ある現象を説明（予測）する

重回帰分析、など。

→アウトプット：係数表、予測スコア

2. 複雑なデータを要約（整理）する

クラスタ分析、など。

→アウトプット：重心表、クラスタ番号

6. 実例：クラスタリング

順番は逆転してしまいましたが、最初にクラスタリング（データを要約）から説明していきましょう。マーケティングの場面では、予測したい現象が明らかであるケースが少ないことが多いからです（あったとしても非常に稀な現象かもしれない）。

1. ある現象を説明（予測）する

重回帰分析、など。

→アウトプット：係数表、予測スコア

2. 複雑なデータを要約（整理）する

クラスター分析、など。

→アウトプット：重心表、クラスター番号

http://daas.la.coocan.jp/GLM/tahenryou_04_cluster.htm

7. 実例：スコアリング

もしもクラスタリングなどで「ターゲット」となるべき対象が見いだせたら、そのターゲットに近い（所属するもしくは類似する）ユーザーを見つけたいものです。見つけることができれば、彼らにアプローチすることが効率的だからです。

1. ある現象を説明（予測）する

重回帰分析、など。

→アウトプット：係数表、予測スコア

http://daas.la.coocan.jp/GLM/5_multiple_regression.htm

2. 複雑なデータを要約（整理）する

クラスター分析、など。

→アウトプット：重心表、クラスター番号

8. なぜデータ分析は役に立たないのか？

データ分析が役に立たないのは、正確に表現すると以下のようなことを指していると思われる。

- ・ **統計学的な考え方**を理解していないため、計算で始終している
- ・ 様々な手法それぞれの**特徴や制約**を知らないため**結果の解釈**ができない
- ・ **固有科学の問題**を統計学の問題と誤ってしまっている

8. なぜデータ分析は役に立たないのか？

また、ビジネスで役立てようとする、統計学とは全く関係のない「仕事の進め方」が邪魔をして存分に専門知識を活かすことができないということもあるでしょう。よくつまづく例として、以下のようなものがあげられます。

- ・お客様（報告相手）との**コミュニケーション**が成立していない
- ・統計学で**効率的に解ける課題**を見誤っている
- ・解析者の**スキルが偏っている**（マネージャーとアナリストの乖離）

8. なぜデータ分析は役に立たないのか？

最近では簡単に分析アウトプットが出せるサービスやツールが多く出てきているため、誰でもデータがあればアウトプットを出せるようになりました。しかしながら、これによって「大切なことを無視して、考えない」ようになってしまっているケースも少なくありません。本来、解析ツールは「考える時間を増やす」ためのものです。

【教訓】

**ひとつの結果が出たら、
その結果を出すために使った時間の2倍も3倍も考えましょう。**

9. 本当に大切なたった一つのこと

統計学をビジネスに活かすために、一つだけ大切なことをあげるとすると

・考えること

としか言いようがありません。

ここに含まれる3つの意図をあえて具体的に書いてしまうと：

1. 統計手法の本質はシンプルであり、手法によってさほど大差ないと理解する
2. とにかくデータをよく見て実際に起きていることを想像すること
3. そのアウトプット出すために何が必要か、出したらどう使えるか自問自答する

9. 本当に大切なたった一つのこと

統計学をビジネスに活かすために、一つだけ大切なことをあげるとすると

- ・考えること

としか言いようがありません。

ここに含まれる3つの意図をあえて具体的に書いてしまうと：

1. 統計手法の本質はシンプルであり、手法によってさほど大差ないと理解する
2. とにかくデータをよく見て実際に起きていることを想像すること
3. そのアウトプット出すために何が必要か、出したらどう使えるか自問自答する

エピソード

私は大学2年生のときに、初めて統計学の本を買って勉強をしました。大学時代は心理学科だったものの、ほとんど統計の勉強しか記憶に残っていないほど勉強をしていました。その後、大学院にまで進学してデータマイニングの専門会社に入社しました。現在に至るまで学術的にも実務的にも、**かれこれ10年以上統計学に携わり続けて**います。その私がこれほどまでに「手法チックな話をしない」のは、やはり**ビジネスにおいて重要なこと**が、それ以上にあるからです。ただし、一方でこのように主張できるのは、**学術的な下積み**時代が長いということの裏返しでもあります。つまり、かなり多くの知識があったうえで、実務経験を積むことでしか**現状は統計学をビジネスに活かすことは「まだまだ難しい」と感じている**ということです。ただ、職人になれということでは決してなく、必要な知識を最小限・効率的に学び、自身のビジネスの中でそれを活かすために**「自分で考える」ということは、絶対に欠かせないこと**だと強く主張したいという思いがあります。もし、自分で考えたいという人がいるならば、その人の手助けをしたいと昔から今に至るまで思い続けています。