



# データマイニング実践ガイド

第2回

# 目次

1. 科学の基本：観測、仮説、説明、予測
2. 演繹と帰納
3. データ：形式、変数型、尺度水準
4. 予測モデルとは（教師あり機械学習）
5. 分析の事前要件設定
6. 正解定義
7. 説明変数のデータ加工
8. モデルの評価
9. ~~モデルの当てはめ（予測）~~
10. 質問紙調査とデータ分析：信頼性と妥当性

イントロダクション

データ分析の準備編

データ分析

エピローグ

# イントロダクション

1. 科学の基本：観測、仮説、説明、予測
2. 演繹と帰納

# 本日、学べること

- 科学的な考え方をもってデータと向き合えるようになる
- モデル解析という基本を理解できるようになる
- データ解析の裏側で何が起きているのか知ることができる

# 1. 科学の基本：観測、仮説、説明、予測

科学とは単一事例的な観測から始まり、ある事象を予測することができて「ヒトの行動を変える」ことに本質的な意義がある。

例)

- 今日は雨が降る→傘を持っていく→ぬれずにすんだ  
※傘を持っていくという行動を起こさなければ意味がない
- 今日は雨が降る→サングラスをしていく→ぬれた  
※無意味な行動は、やはり無意味

# 1. 科学の基本：観測、仮説、説明、予測

精度よく予測できることは重要

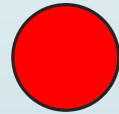
ただし

- ・ **適切な行動を起こせない予測**

は何も変えることができない。

# 1. 科学の基本：観測、仮説、説明、予測

科学の基本である観測、仮説、説明、予測を体感するためのゲーム  
を試してみる。

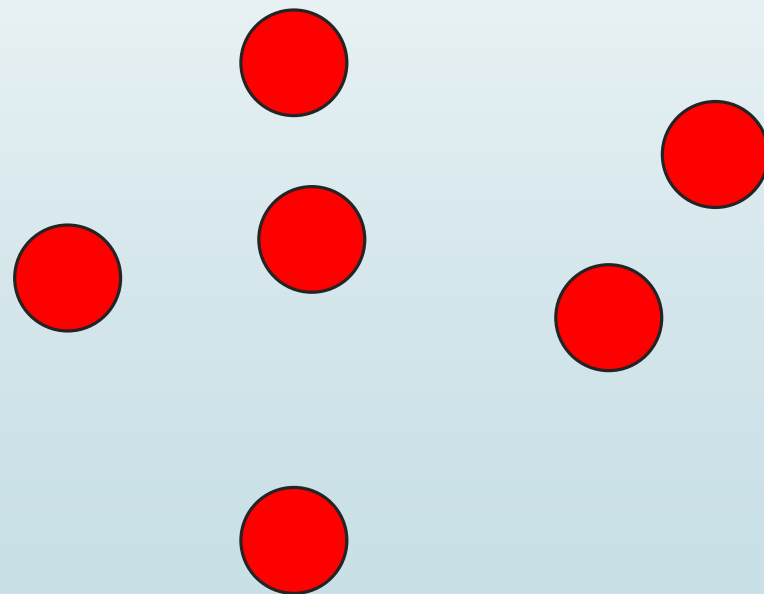


**赤い丸だ**

※形状と色のことをいっている。

# 1. 科学の基本：観測、仮説、説明、予測

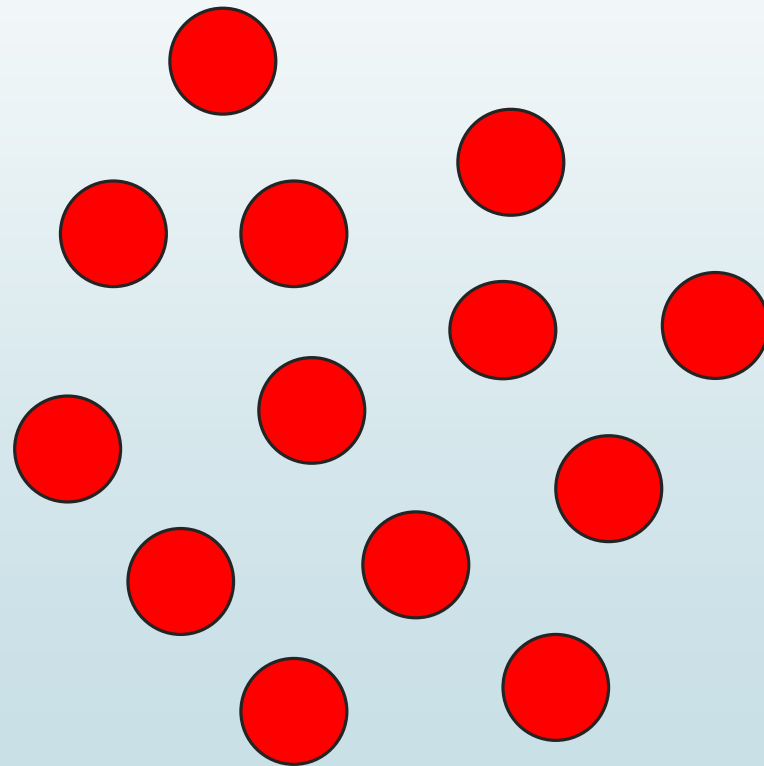
科学の基本である観測、仮説、説明、予測を体感するためのゲーム  
を試してみる。



丸は赤色**だろう**

# 1. 科学の基本：観測、仮説、説明、予測

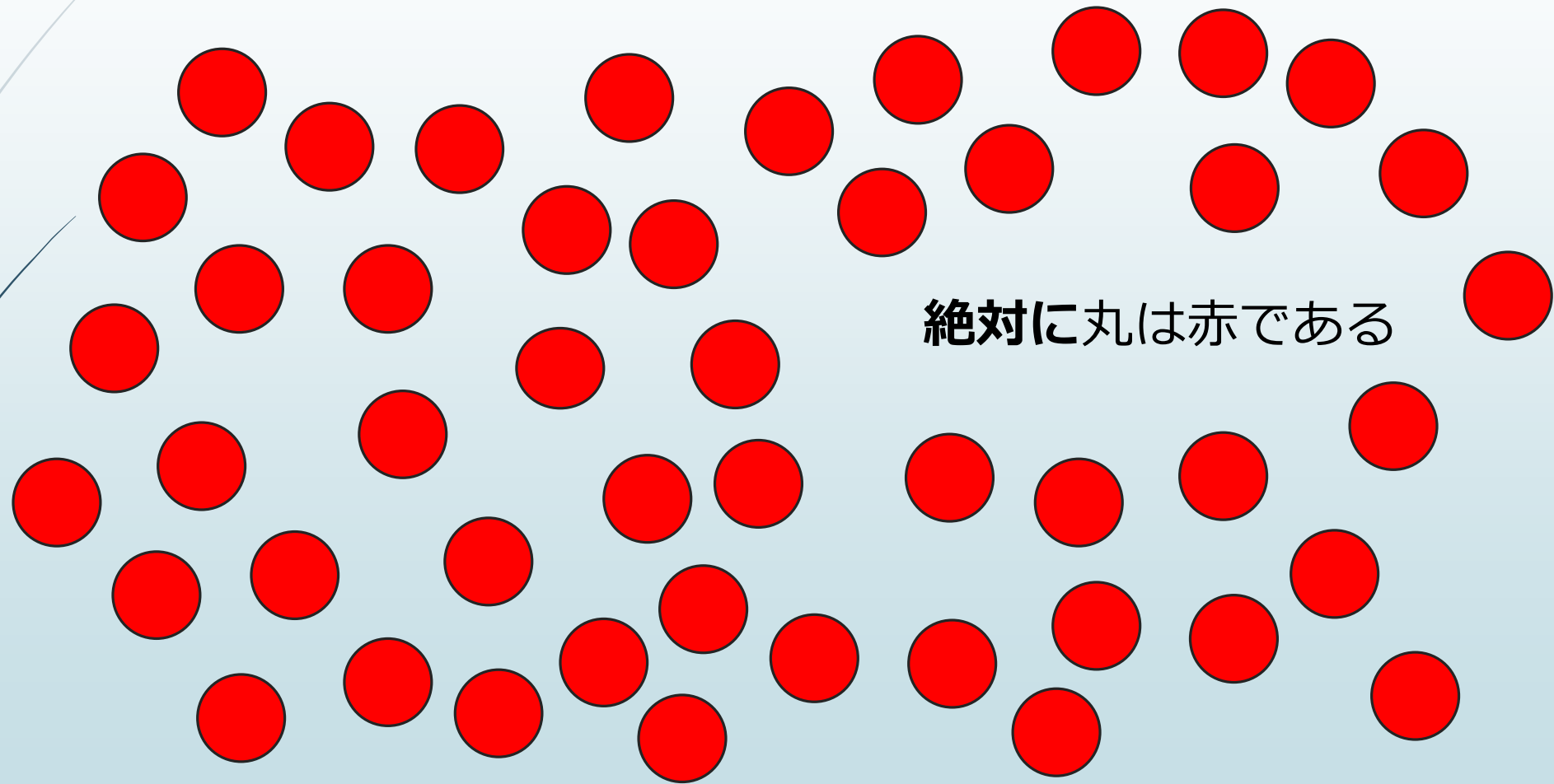
科学の基本である観測、仮説、説明、予測を体感するためのゲーム  
を試してみる。



**ほぼ間違いなく丸は赤である**

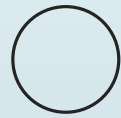
# 1. 科学の基本：観測、仮説、説明、予測

科学の基本である観測、仮説、説明、予測を体感するためのゲーム  
を試してみる。



# 1. 科学の基本：観測、仮説、説明、予測

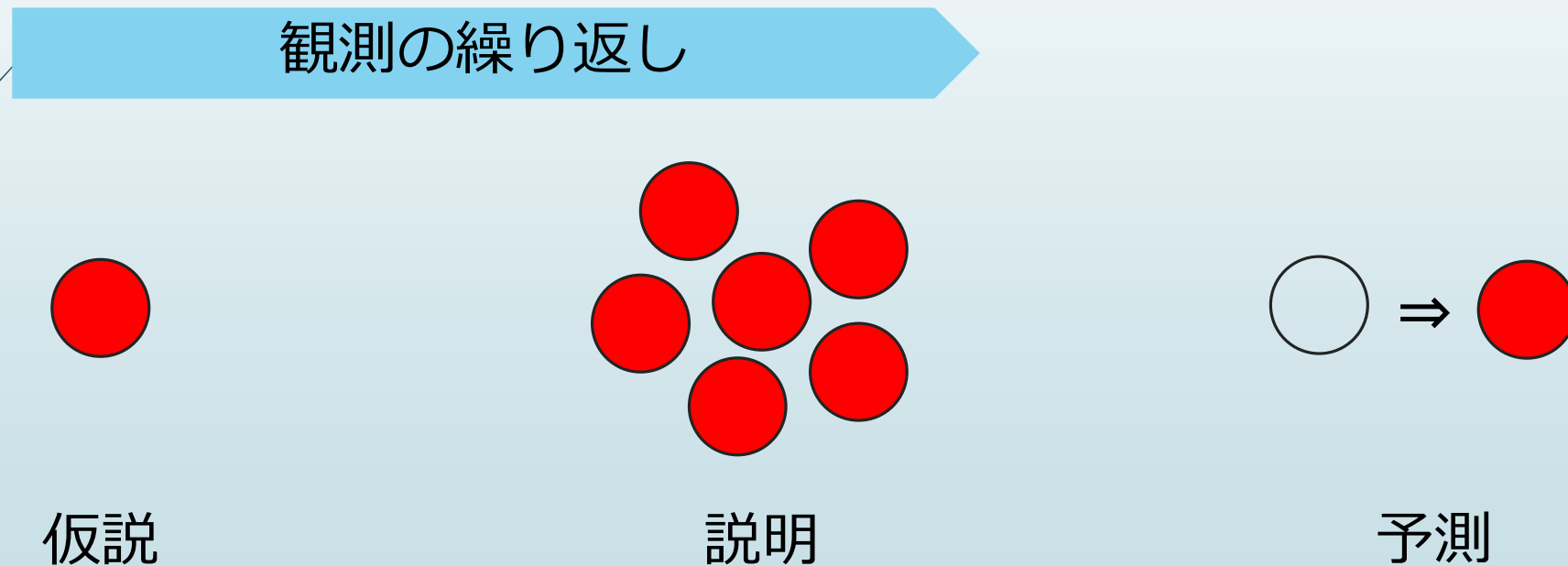
科学の基本である観測、仮説、説明、予測を体感するためのゲーム  
を試してみる。



さて、丸は何色？

# 1. 科学の基本：観測、仮説、説明、予測

ある事象が観測されることで「丸は赤ではないか？」という仮説が生まれ、その事象が繰り返し観測され、それをモデル化（説明⇨一般化）することで、予測ができるようになる。



## 2. 帰納と演繹

### ■ 帰納（きのう）

物事の積み重ねから全てをいうようなアプローチ

- ・ d1: 丸は赤である, d2: 丸は赤である  
→丸は赤である（に違いない）

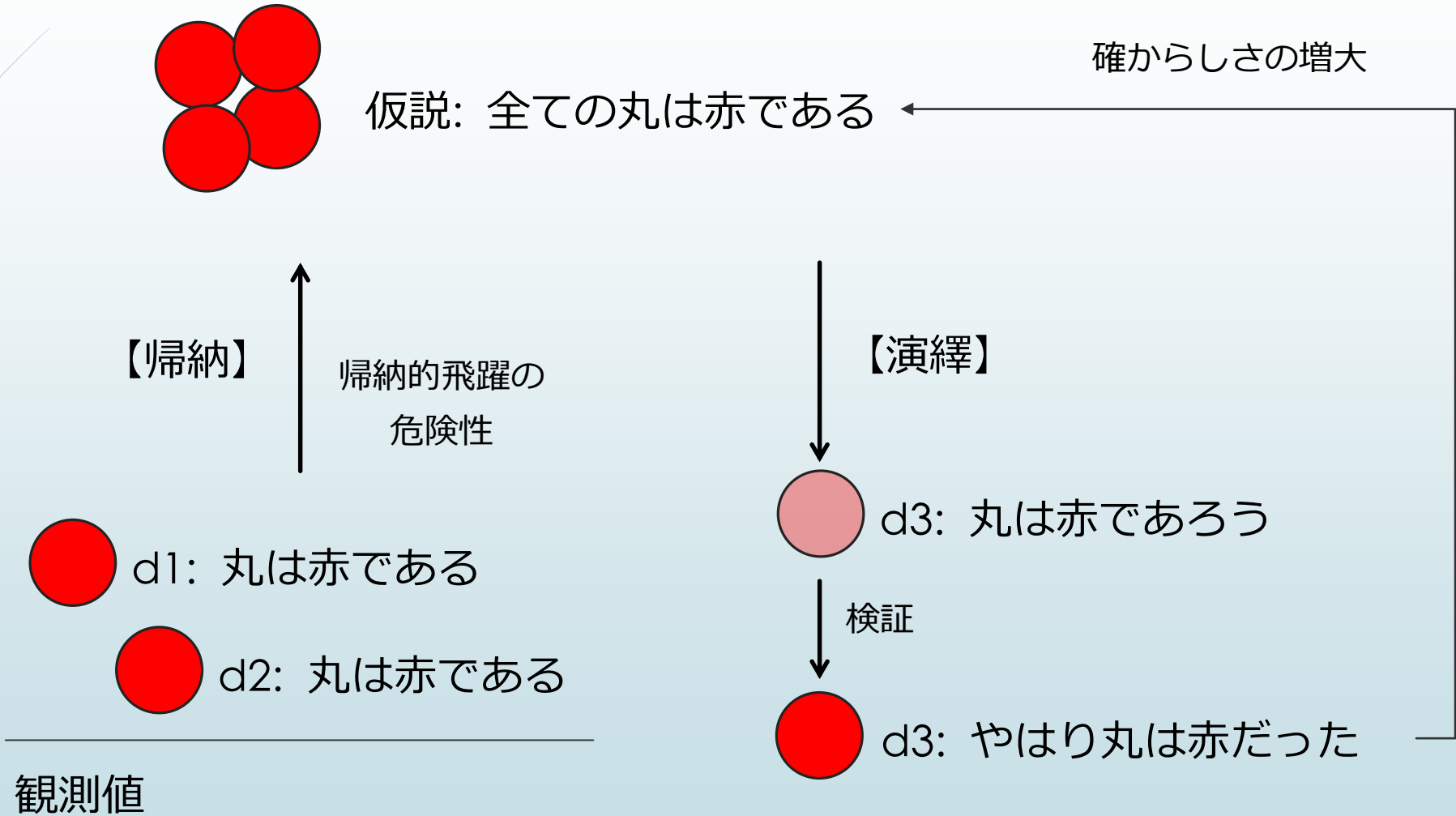
### ■ 演繹（えんえき）

全てのことを言いきってから、ある物事を改めて言い直す

- ・ 全ての丸は赤である（であろう）  
→d3: やはり丸は赤であった

※説明の便宜上、正確でない表現をしています。

## 2. 帰納と演繹



## 2. 帰納と演繹

無意識に「やれてしまっている」日々の行動：

1. 観測： バナーを小さくした方がCVあがるんじゃないか？

↓ 帰納

2. 仮説： きっとバナーを小さくした方がCVが上がるはず。

3. 説明： 分析した結果、やはり間違いない（モデル化⇔一般化）

↓ 演繹

4. 予測： 別の施策でも試してみたらCVあがった（確からしさの増大）

# データ分析の準備編

3. データ：形式、変数型、尺度水準
4. 予測モデルとは（教師あり機械学習）
5. 分析の事前要件設定
6. 正解定義
7. 説明変数のデータ加工

### 3. データ：形式、変数型、尺度水準

データといっても形式、変数型、さらに細かくいえば尺度水準を理解していなければなりません。これらを区別できれば、後に出てくるデータ加工がスムーズになります。作業をしない人にとっても重要な概念です。

■ **形式**： マスタ形式か？ トランザクション形式か？

■ **変数型**： 連続型か？ カテゴリカル型か？

■ **尺度水準**： 名義、順序、間隔、比率

[http://daas.la.cocacn.jp/jisen\\_datasci/jds\\_01\\_data.htm](http://daas.la.cocacn.jp/jisen_datasci/jds_01_data.htm)

## 4. 予測モデルとは（教師あり機械学習）

「**教師あり機械学習**」とは、予測したい事象が明確である場合の分析手法です。これに対して予測したい事象が明確でない場合は「**教師なし機械学習**」というものを使います。

この分析を適用する際には以下の質問に答えられることが大切です：

- **目的変数** 何を予測したいのか？ それは明文化できるもので、かつデータとして表現できるものであるか？
- **説明変数** 予測したい事象を説明できそうな要因は何か？ 以下同文。

[http://daas.la.coocan.jp/jisen\\_datasci/jds\\_03\\_toukei.htm](http://daas.la.coocan.jp/jisen_datasci/jds_03_toukei.htm)

## 5. 分析の事前要件設定

ほとんどの場合において 分析＝モデル作成 という文脈で使われていますが、分析をすることそのものに大した技術は必要ありません（ツールを使えば誰でもできます）。

当たり前のように聞こえるかもしれませんが、**何を分析するのか明確にできていない（できていると思っ**てしまっている）ケースはよくあります。

データがあれば分析できるというのは間違いで、何を分析するかによって必要なデータが決まってくるものです。分析の仕方が分からないというケースでは、「仕方」ではなく「何をしたいか」が分かっていないことが多いです。

## 5. 分析の事前要件設定

より具体的にモデルと予測の区別をしてみましょう。

■ **モデル** モデルを作るということは、ある予測したい事象を「説明する」ということです。

[http://daas.la.coocan.jp/jisen\\_datasci/jds\\_06\\_setsumei.htm](http://daas.la.coocan.jp/jisen_datasci/jds_06_setsumei.htm)

■ **予測** 予測とは、モデルを当てはめることで予測したい事象を「予測する」ということです。

[http://daas.la.coocan.jp/jisen\\_datasci/jds\\_07\\_yosoku.htm](http://daas.la.coocan.jp/jisen_datasci/jds_07_yosoku.htm)

## 5. 分析の事前要件設定

例えば、あるwebサイトの中から、購買しやすいユーザーを見つけたいというケースを考えたとき、**モデル作成対象者を明文化**できるでしょうか。



**来訪者全員**：とにかくwebサイトに来訪している人すべて。ある施策を打つべき対象は多くの場合、一部であるはず。

**施策対象候補者**：理論上、ある施策を打つ対象として考えられるユーザー。

**モデル作成対象者**：実際にデータが取れていて分析可能なユーザー。

## 5. 分析の事前要件設定

こうした事態に遭遇したとき、以下の質問がでるようではなくてはなりません。

### 1) 購買しやすいユーザーとは具体的にどういう人ですか？

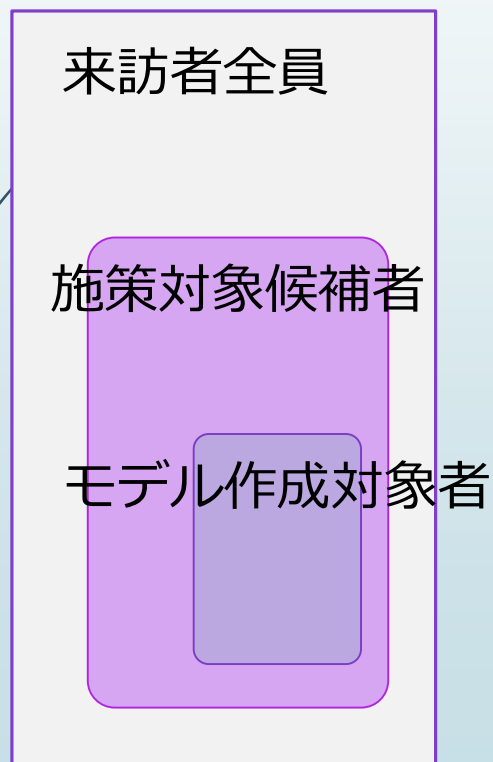
2018.7.1～7.10のバーゲン中にメルマガ経由で流入して、7.20までに5万円以上のバッグを1店以上購買したユーザー。 →購買しやすいとは、実はキャンペーンに反応しやすいということだった。

### 2) そのユーザーを見つけられたらどういう効果が見込めますか？

2018.9.1～9.10に同様のバーゲンをするので、購買しやすい人にだけ限定でメルマガを送りたい。その際に特別展示会場の案内を送るため、対象を絞ることで展示会場の客質を高めたい。 →真の目的は新作の売上をあげることであった。

## 5. 分析の事前要件設定

例えば、あるwebサイトの中から、購買しやすいユーザーを見つけたいというケースを考えたとき、**モデル作成対象者を明文化**できるでしょうか。



**施策対象候補者**：次回のメルマガ配信可能な人すべて。

**モデル作成対象者**：今回のメルマガが配信された人で、購買履歴など分析に必要なデータがある人（この中で「買いやすい人」を正解にしてモデルをつくる）。

※なおここに時間の概念の説明は省略してある。  
難しくなるので割愛。

## 5. 分析の事前要件設定

分析（モデル作成）するときのマジックボックス：

- 1) 正解の定義、何を予測したいのか？
- 1) ' それを使って何がしたいのか？
- 2) 対象は誰か？ ※誰に当てはめるのか？
- 3) どんな要因で説明できそうか？

ID	X1, X2, X3, ..., X_p	Y
2)	3)	1) + 1) '

## 6. 正解定義

正解の定義で必要かつ重要なこと：

- ・まず具体的に明文化すること → ビジネス側の責任
- ・そのうえでデータとしてどう表現できるかを考えること → 分析側の責任
- ・表現できないことは代替案で妥協すること → ビジネス側の責任

### 【教訓】

簡単に分析なんてできません。

たかが正解ひとつ決めることさえ普通の人にはできません。

## 【再掲】 5. 分析の事前要件設定

こうした事態に遭遇したとき、以下の質問ができるようになってはなりません。

### 1) 購買しやすいユーザーとは具体的にどういう人ですか？

**2018.7.1～7.10のバーゲン中にメルマガ経由で流入して、7.20までに5万円以上のバッグを1店以上購買したユーザー。** →購買しやすいとは、実はキャンペーンに反応しやすいということだった。

### 2) そのユーザーを見つけられたらどういう効果が見込めますか？

2018.9.1～9.10に同様のバーゲンをするので、購買しやすい人にだけ限定でメルマガを送りたい。その際に特別展示会場の案内を送るため、対象を絞ることで展示会場の客質を高めたい。 →真の目的は新作の売上をあげることであった。

## 7. 説明変数のデータ加工

さて、正解が決まり、分析対象も決めることができたとして、次はデータ加工をしなければなりません。データ加工とは加工そのものの技術も必要ですが、それ以前に「データで表現する」という独特な技術が必要です。ここでは加工そのものの技術を紹介します（後者は最後に紹介します）。

## 【再掲】 3. データ：形式、変数型、尺度水準

データといっても形式、変数型、さらに細かくいえば尺度水準を理解していなければなりません。これらを区別できれば、後に出てくるデータ加工がスムーズになります。作業をしない人にとっても重要な概念です。

■ **形式**： マスタ形式か？ トランザクション形式か？

■ **変数型**： 連続型か？ カテゴリカル型か？

■ **尺度水準**： 名義、順序、間隔、比率

[http://daas.la.cocacn.jp/jisen\\_datasci/jds\\_01\\_data.htm](http://daas.la.cocacn.jp/jisen_datasci/jds_01_data.htm)

# 7. 説明変数のデータ加工

データ加工をするといっても、いきなりできるわけでもなく、どんなデータがあるのかを整理し、どのデータをどう加工するかの仕様を決めてから作業開始となります。

## ■ データ整理

[http://daas.la.coocan.jp/jisen\\_datasci/jds\\_04\\_dataseiri.htm](http://daas.la.coocan.jp/jisen_datasci/jds_04_dataseiri.htm)

## ■ データ加工

[http://daas.la.coocan.jp/jisen\\_datasci/jds\\_05\\_datakakou.htm](http://daas.la.coocan.jp/jisen_datasci/jds_05_datakakou.htm)

# データ分析

- 8. モデルの評価
- 9. モデルの当てはめ（予測）

## 8. モデルの評価

データ加工が完了して、モデル作成用のデータができれば「分析」そのものは大したことはありません（ちょっと誤解を招く表現ですが、あえて）。なぜなら、いわゆる「教師あり機械学習」というものは、目的変数の「型」によって使える手法が決まっているからです。いくつかの手法の選択肢はありますが、大差ありません（これも誤解を招く表現ですが、あえて）。

■ 目的変数 = **連続型** ⇒ 重回帰分析

■ 目的変数 = **カテゴリカル型** ⇒ ロジスティック回帰分析

[http://daas.la.coocan.jp/GLM/1\\_modelkaiseiki\\_kangaekata.htm](http://daas.la.coocan.jp/GLM/1_modelkaiseiki_kangaekata.htm)

## 8. モデルの評価

細かくいえば、採用した手法によりモデル精度の評価の仕方は違います。しかし、本質的には以下の2点について考えることが重要です。

### 1) 統計学的な観点での評価

これは予測と実績が一致しているか？ということです。

### 2) ビジネス的な観点での評価

これは作成されたモデルが妥当か？ということです。

## 8. モデルの評価

モデルの精度とは**予測と実績が一致しているか？**ということです。

目的変数 = 連続型

ID	予測値	実績値
1	10.5	10
2	15.2	10
3	9.0	9
4	6.5	5
5	11.1	12

ID=2はアヤシイが、ほぼうまく予測できている気がする。

目的変数 = カテゴリカル型

ID	予測値	実績値
1	0.501	0
2	0.005	0
3	0.781	1
4	0.665	1
5	0.209	0

ID=1はアヤシイが、ほぼうまく予測できている気がする。

## 8. モデルの評価

目的変数が連続型の場合は、想像しやすい。

目的変数 = 連続型

ID	予測値	実績値	残差
1	10.5	10	-0.5
2	15.2	10	-5.2
3	9.0	9	0.0
4	6.5	5	-0.5
5	11.1	12	0.9

※予測値と実績値の差を書いている。



予測値と実績値の差（残差）  
がどれも小さそうなので、い  
いモデルっぽい気がする。

※実際には「決定係数」という  
指標で予測精度を評価します。

## 8. モデルの評価

目的変数がカテゴリカル型の場合は、ちょっと想像しにくい。

目的変数 = カテゴリカル型

ID	予測値	実績値
1	0.501	0
2	0.005	0
3	0.781	1
4	0.665	1
5	0.209	0



? ? ?

## 8. モデルの評価

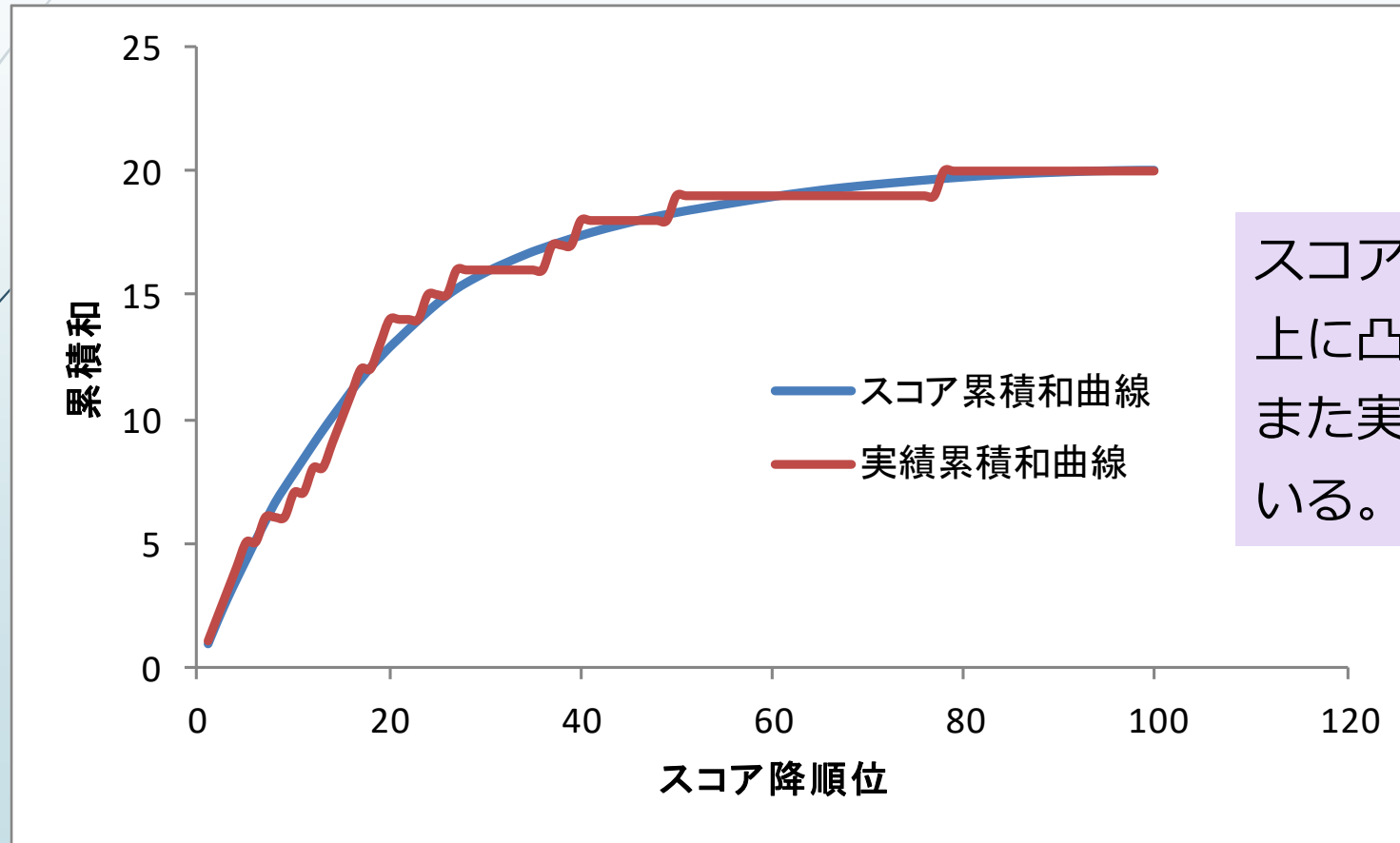
目的変数がカテゴリカル型の場合は、**ゲインチャート**と呼ばれるグラフを描くのがおすすめです。このグラフは単に**予測精度を目視的に理解**できるだけでなく、**ビジネス上使えるモデルになっているか**どうかの示唆も与えてくれます。

▼ゲインチャートを描くためのデータ

ID	予測値	実績値	累積和(予測)	累積和(実績)
3	0.781	1	0.781	1
4	0.665	1	1.446	2
1	0.501	0	1.947	2
5	0.209	0	2.156	2
2	0.005	0	2.161	2

## 8. モデルの評価

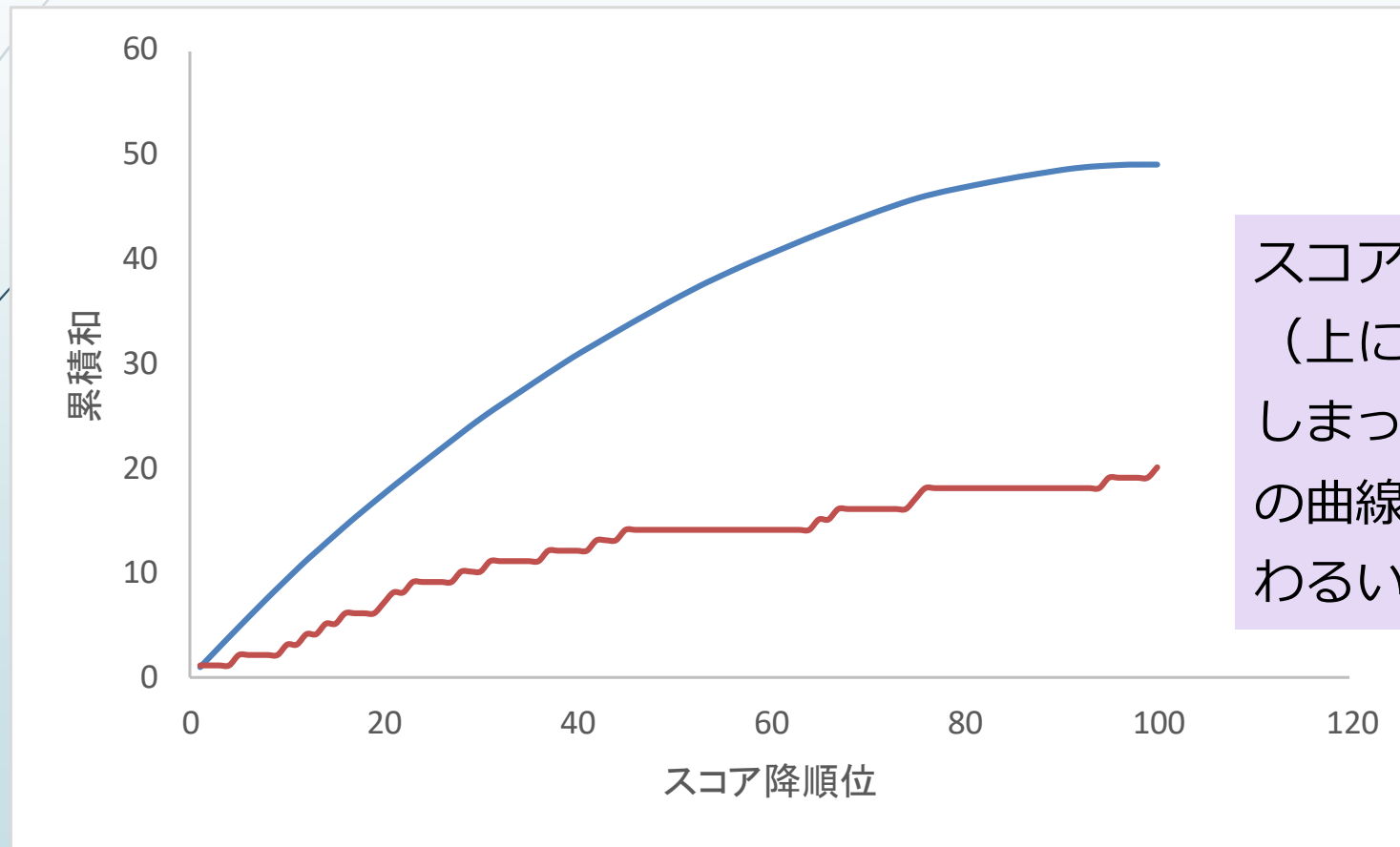
ゲインチャートはROC曲線と呼ばれるものがあり、これと同義のものです。これの本質的な意味を理解するための例示をします。



スコア累積和曲線が滑らかに上に凸の形状を描いている。また実績の曲線とも一致している。≡いいモデル

## 8. モデルの評価

ために先のゲインチャートとして描かれていたスコアを「でたらめなスコア」にして描いてみるとどうなるでしょうか。

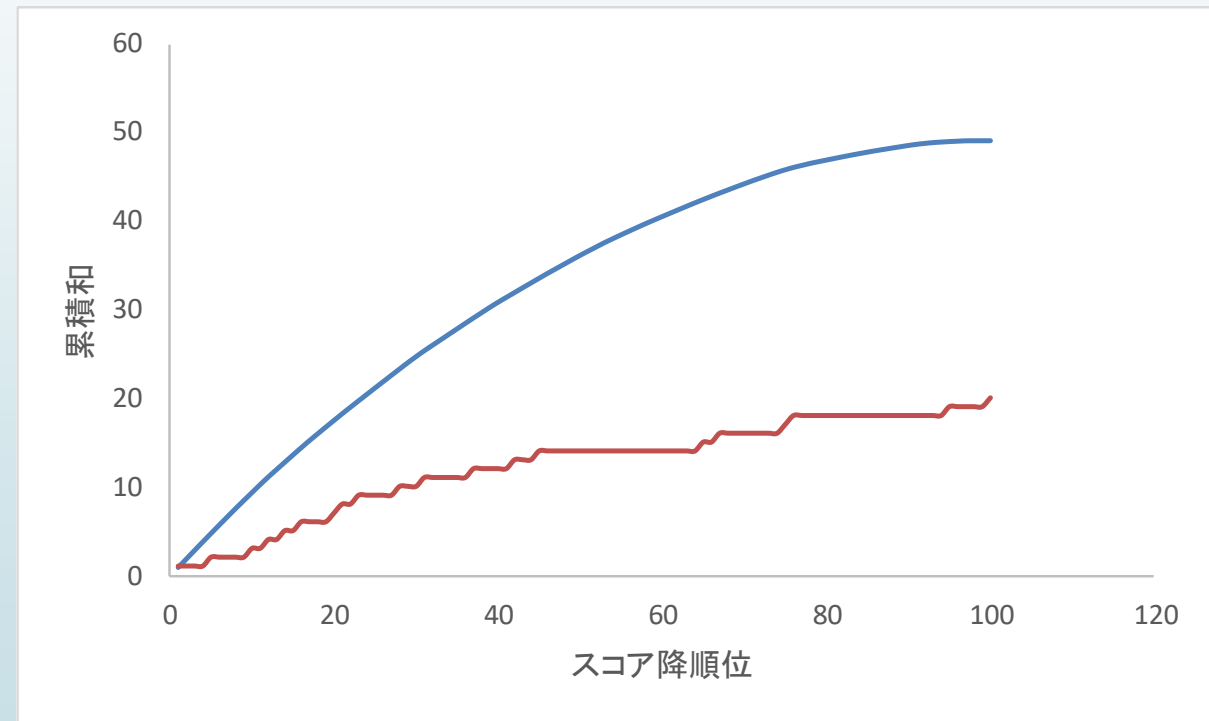
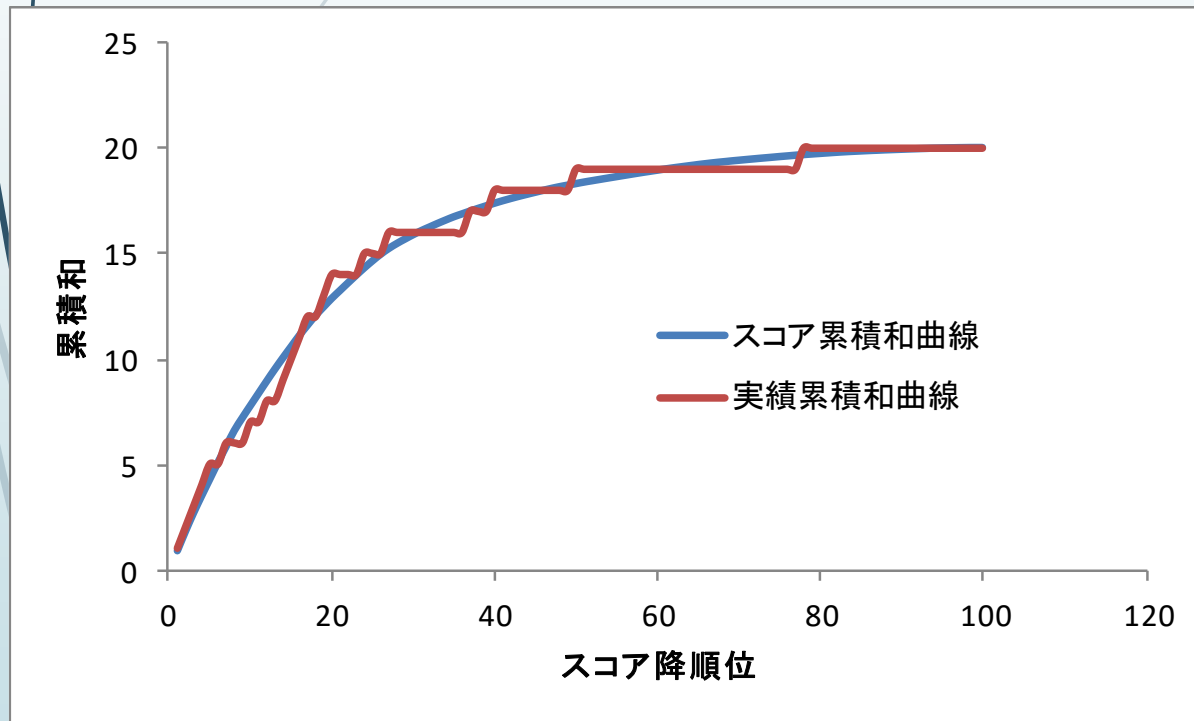


スコア累積和曲線は直線的に（上に凸にならずに）なってしまっている。そのうえ実績の曲線とも乖離している。≠わるいモデル

## 8. モデルの評価

右図：モデルによる予測スコア

左図：ランダムスコア ※ユーザーに単なる乱数を付与しただけ

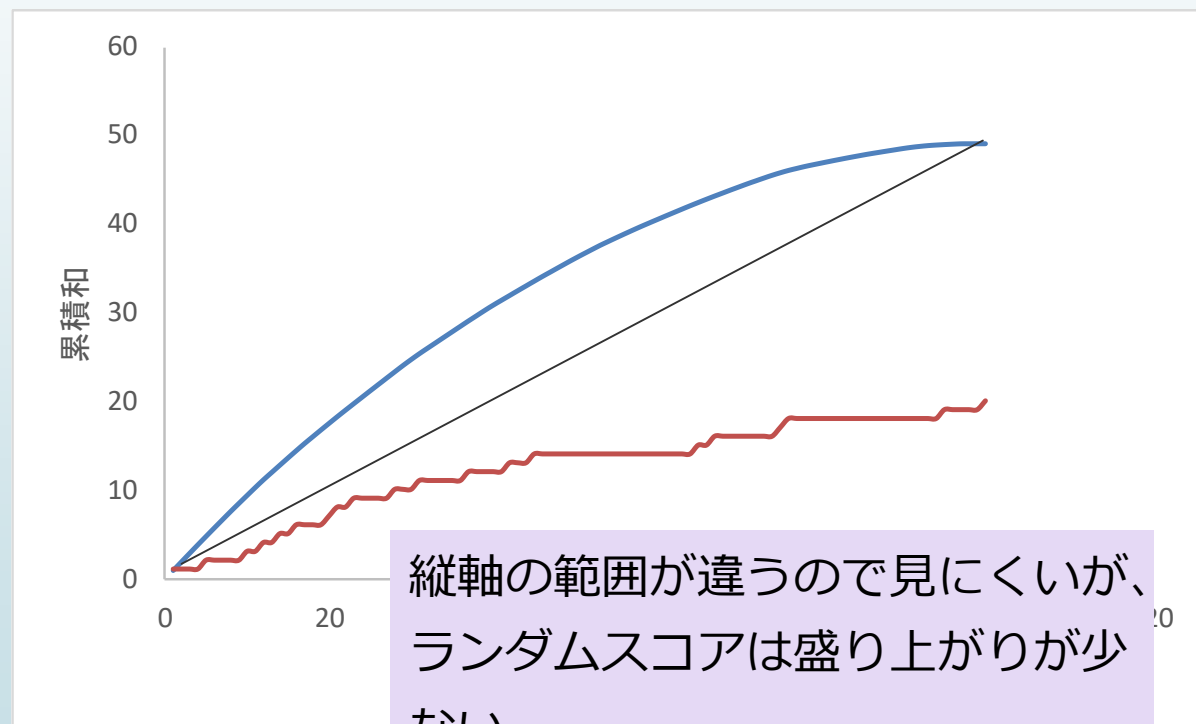
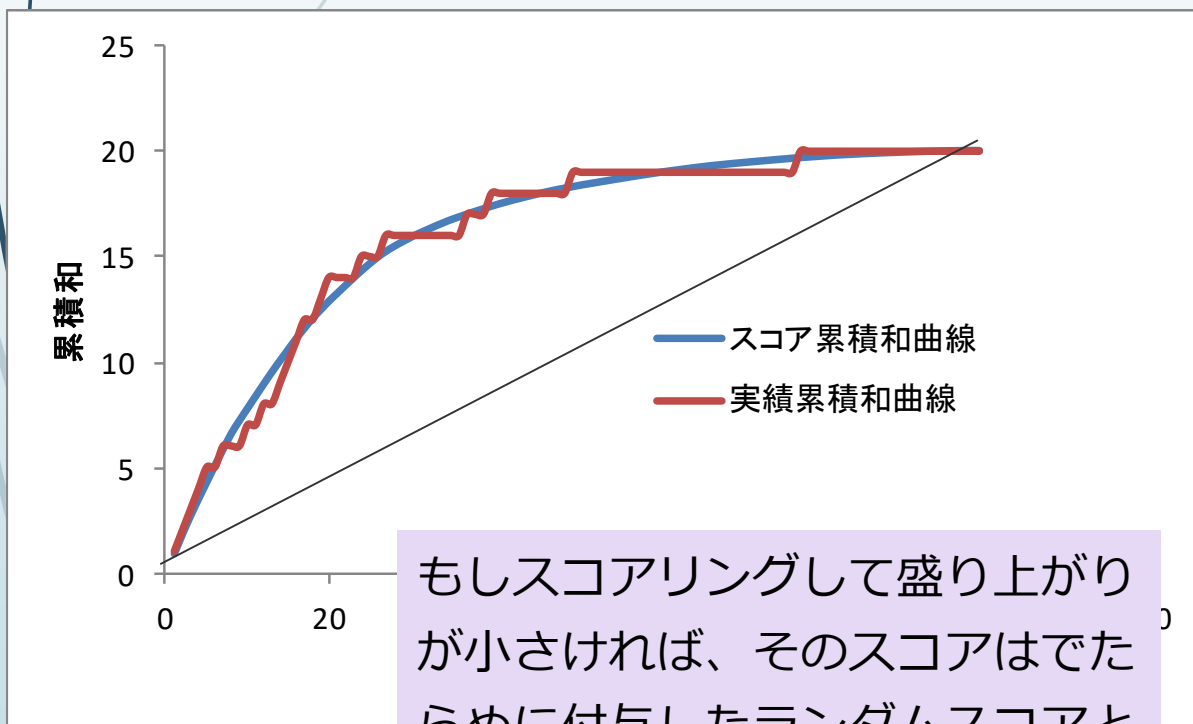


# 8. モデルの評価

## ポイント1

ゲインチャートの盛り上がり小さい、とは何を意味するか？

→でたらめなスコア付けしているのと大差ないということ。



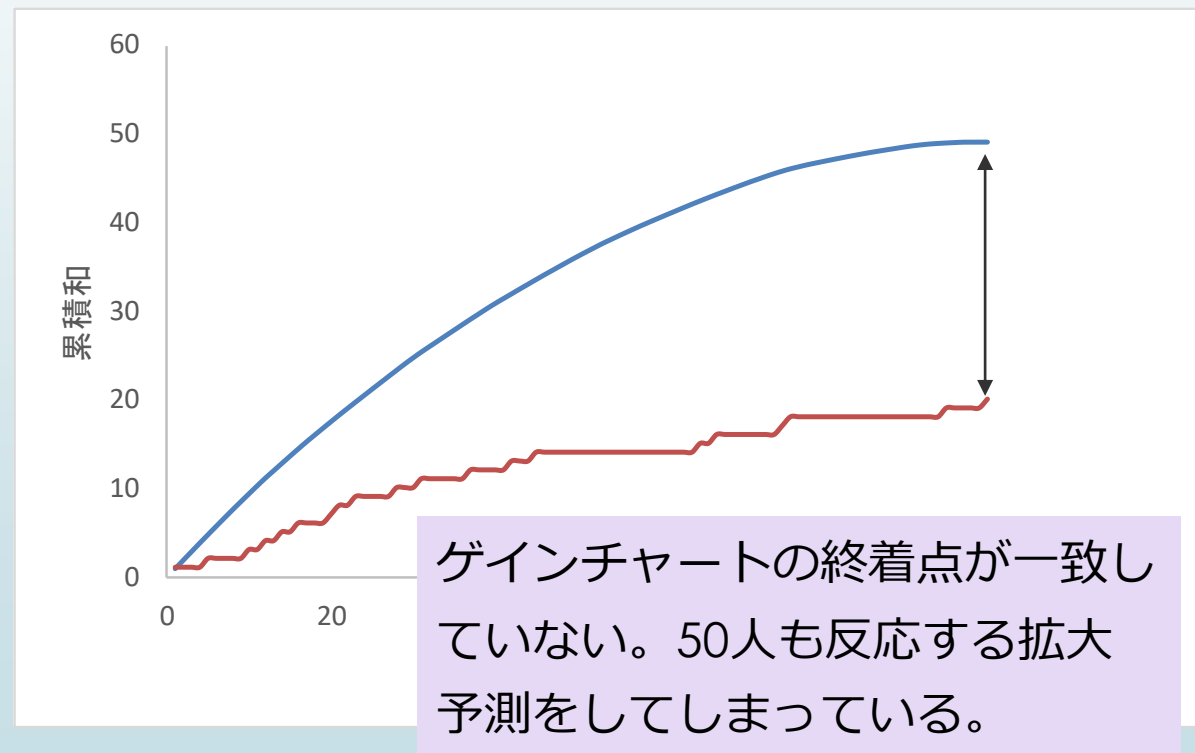
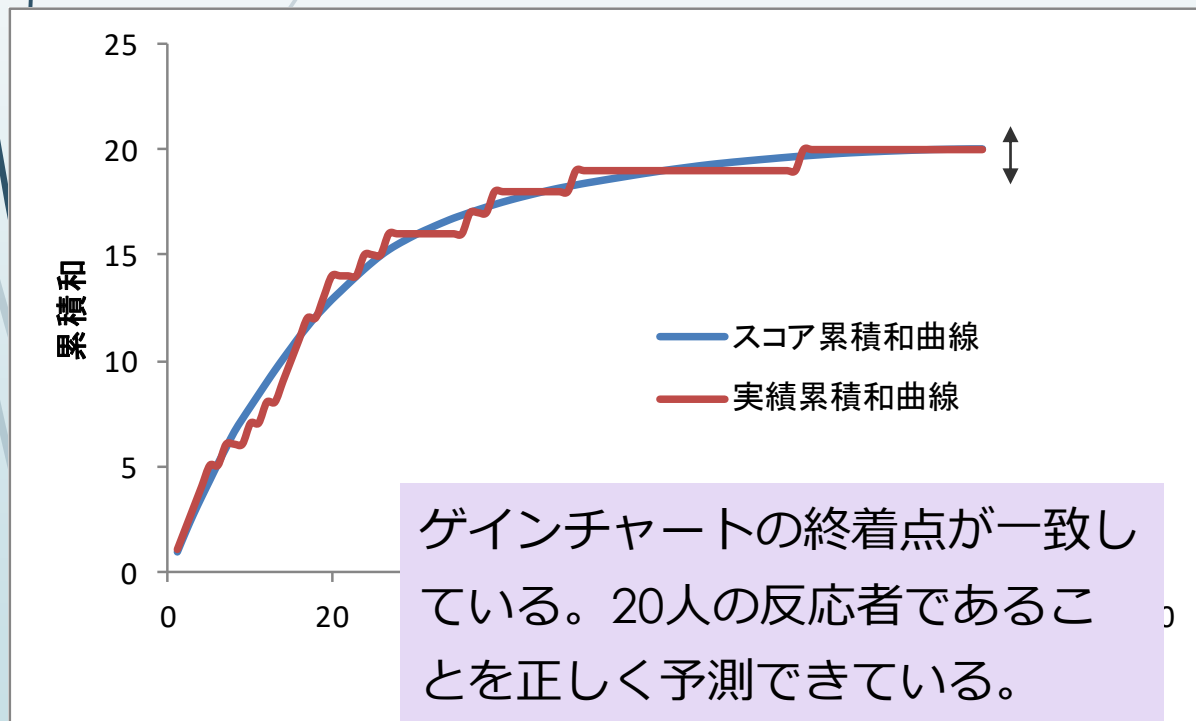
## 8. モデルの評価

### ポイント2

実績のチャートとのずれは何を意味しているか？

→例えば左図の例では、最終的な「反応数」が大きくちがっている

(実際には20人くらいしか反応しないのに50人も反応する予測になってしまっている。)

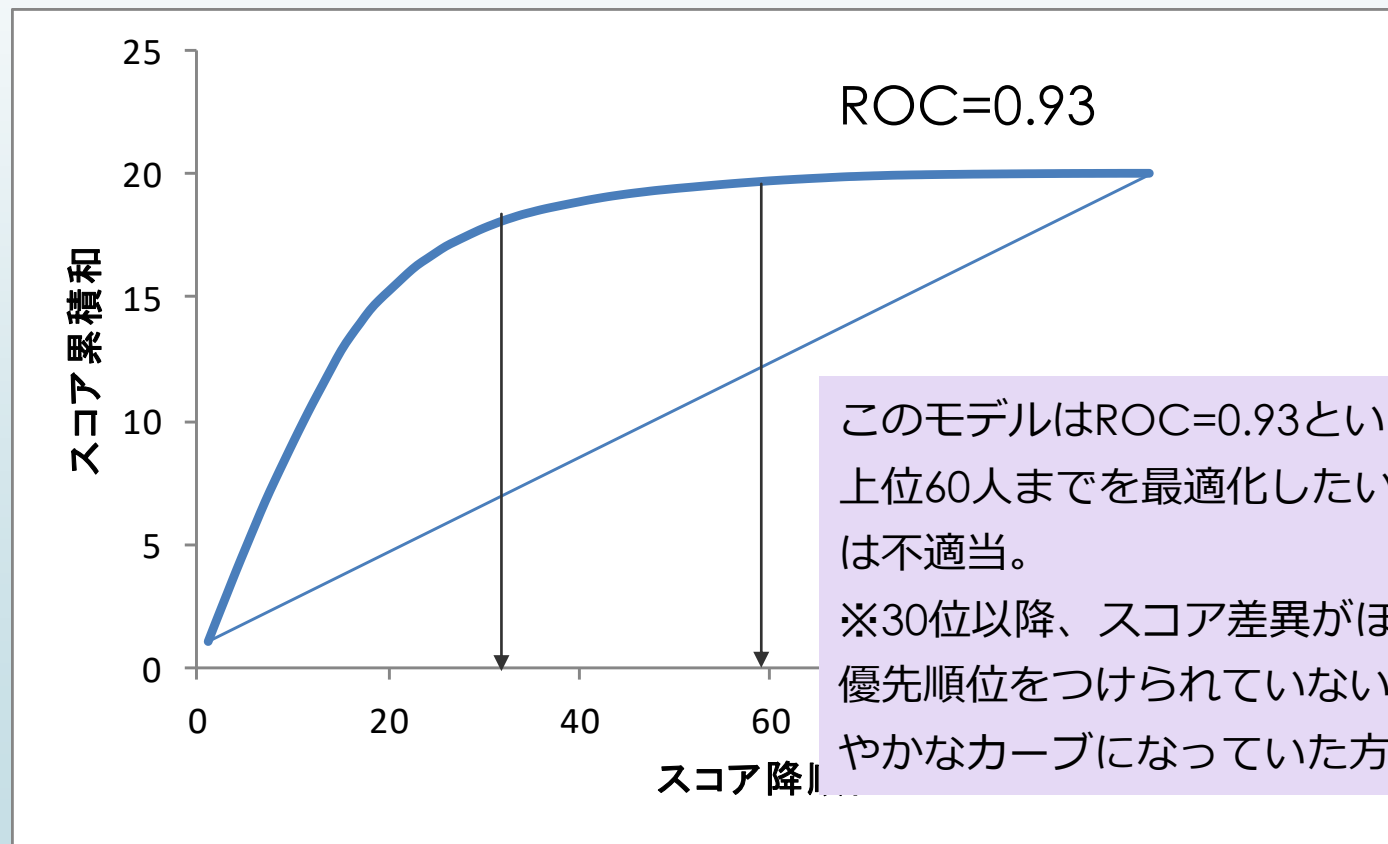


## 8. モデルの評価

### ポイント3

数値的に精度の高いモデルはいいもできるか？

→適用シーンによっては精度指標が高くても使えないことがある。



## ~~9. モデルの当てはめ（予測）~~

いっぺんに説明しても難しくて頭がパンクしてしまうと思うので、今回はモデル作成のところまでとします。

# エピローグ

## 10. 質問紙調査とデータ分析：信頼性と妥当性

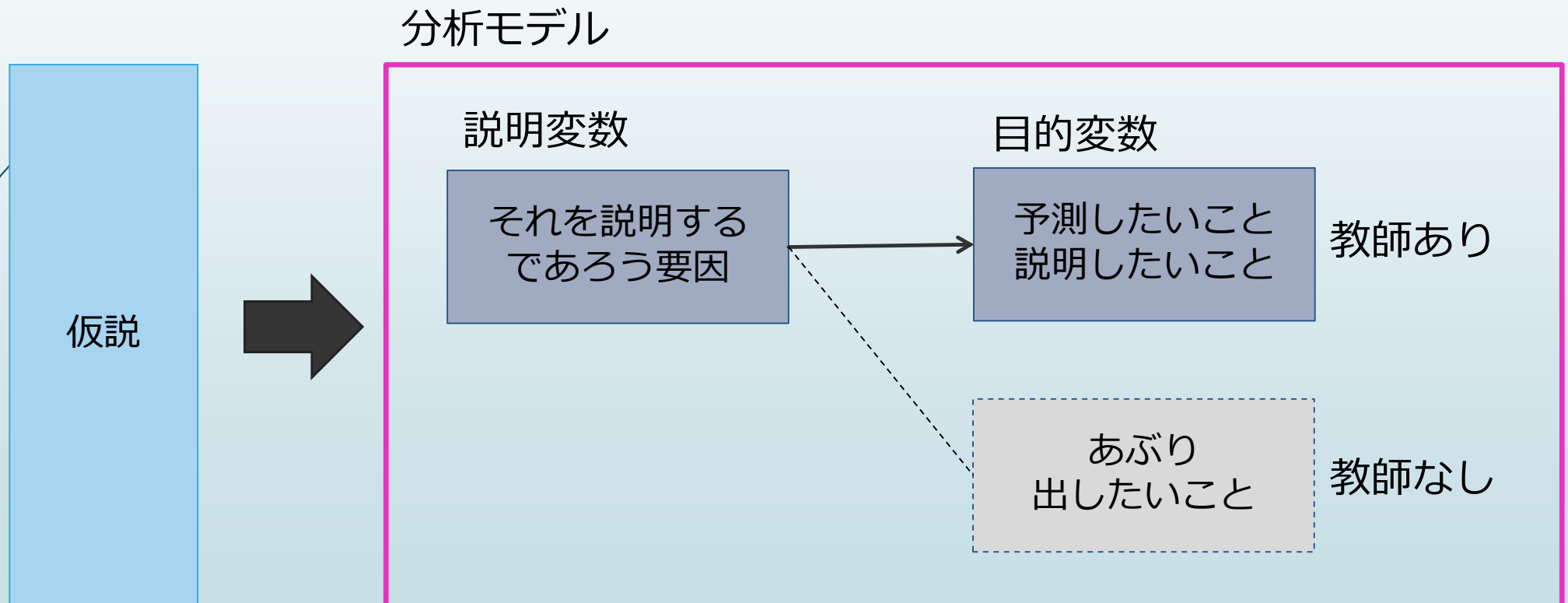
# 10. 質問紙調査とデータ分析：信頼性と妥当性

ここまでのおさらい

- モデル作成には事前の「取り決め」なければならないことがある
  - 正解定義、対象の選定、説明変数の加工仕様
- 何を説明したいかによって分析手法はおのずと決まる
  - ※データがあるから何ができるか？ではない
- 分析は簡単、分析の設定が難しい

# 10. 質問紙調査とデータ分析：信頼性と妥当性

多くの文脈で「アンケート」と言われているものは、正しくは「質問紙調査」といいます。質問紙の設計は本来「非常に難しく高度な技術が必要」となるものですが、もっぱらテキトーにやられてしまっているケースが散見されます。



# 10. 質問紙調査とデータ分析：信頼性と妥当性

質問紙調査を行う際、以下の質問に対して具体的に答えられるか？が大切です。  
以下は若干、極端な例：

- ・何を説明したいのか？

新キャラクターの好意度が、購買意欲につながる・・・×

購買意欲を目的変数としたとき、新キャラクターの好意度を説明変数にすることで、そのモデルは確からしいといえるものになるか・・・○

## 【教訓】

具体的な記述ができないものは、データをとっても検証できない。

# 10. 質問紙調査とデータ分析：信頼性と妥当性

ところで、フツーに使ってしまいがちですが「好意度」とは何のことでしょうか？これも具体的に記述できなければなりません。好意というものの概念は人によってさまざまでしょう。。。

Aさん → 好意とは信頼があるということだ

Bさん → 好意とは無償の愛を捧げられることだ

Cさん → 好意とは嫌いでないことだ

で、結局「好意」ってなんのことでしょうか。  
購買意欲とかも同じですね。

# 10. 質問紙調査とデータ分析：信頼性と妥当性

分析したいモデルが明らかになれば、必要なデータが明らかになります、必要データが明らかになれば、その測定方法が明らかになります。ただし、モノによっては測定する尺度が存在しない、あるいは信頼性や妥当性に乏しいものかもしれません。

- ・ 信頼性と妥当性

[http://daas.la.coocan.jp/toukei\\_hosoku/shinrai\\_and\\_datousei\\_kentou.htm](http://daas.la.coocan.jp/toukei_hosoku/shinrai_and_datousei_kentou.htm)