

目で見てわかるビジネス統計学 ～Excel実践編～

第1回

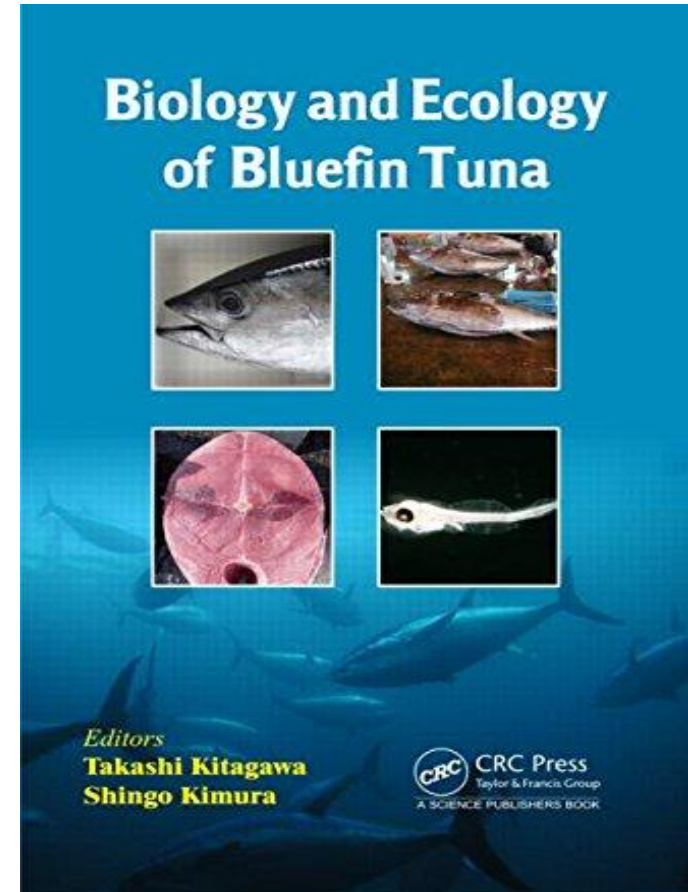
「データの要約と可視化」



和から株式会社

講師紹介

- **氏名**
 - (博士) 門田 実 (かどた みのる)
- **学歴**
 - 北海道大学, University of Rhode Island
学士 (理学)
 - University of Rhode Island, New York
University 修士 (数学・物理)
 - New York University, Columbia University
博士 (応用数学)
- **職歴**
 - New York University 数学講師
 - IPRC(国際太平洋研究センター)
気候変動のモデリング、中期気候予測
 - 近畿大学
21世紀GCOEプログラム研究員 (農学部)
 - Temple University, Japan 准教授 (経済学部)
 - LINE Corp Data Scientist
 - 首都大学 物理非常勤講師
 - 東洋大学 数学非常勤講師



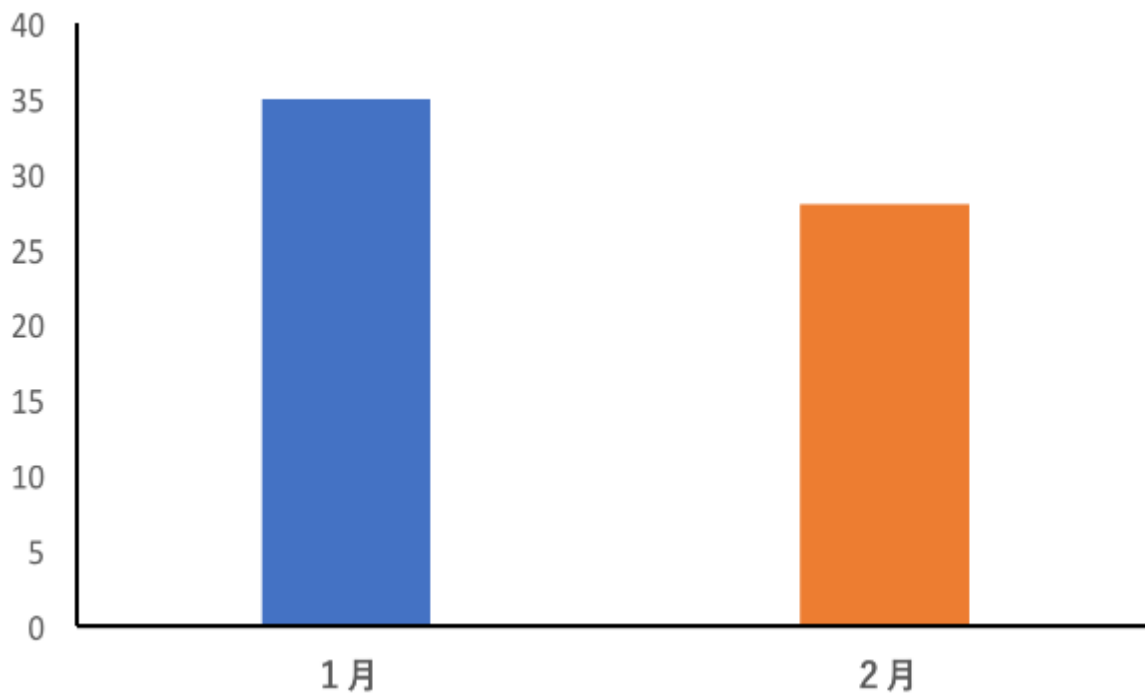
- **趣味**
 - ウルトラマラソン (サハラ250km横断)
 - ブラジリアン柔術

ビジネスデータと統計学

うちの社では月末会議で、その月の
数字集計を行い、全員で課題点を確認し、
共有しているので、統計学など
難しいことは必要ないのでは？



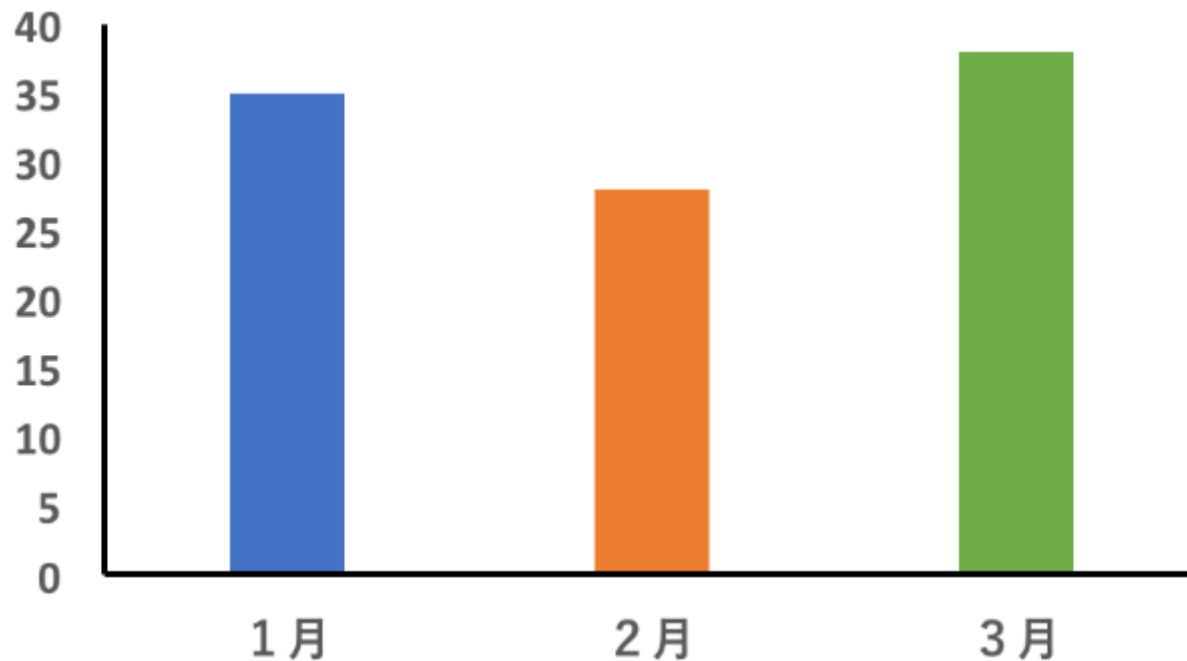
ある月の売上



2月の売上が1月の数字を下回ったので、来月は巻き返しを狙わないといけない。そこでキャンペーンを打とう！



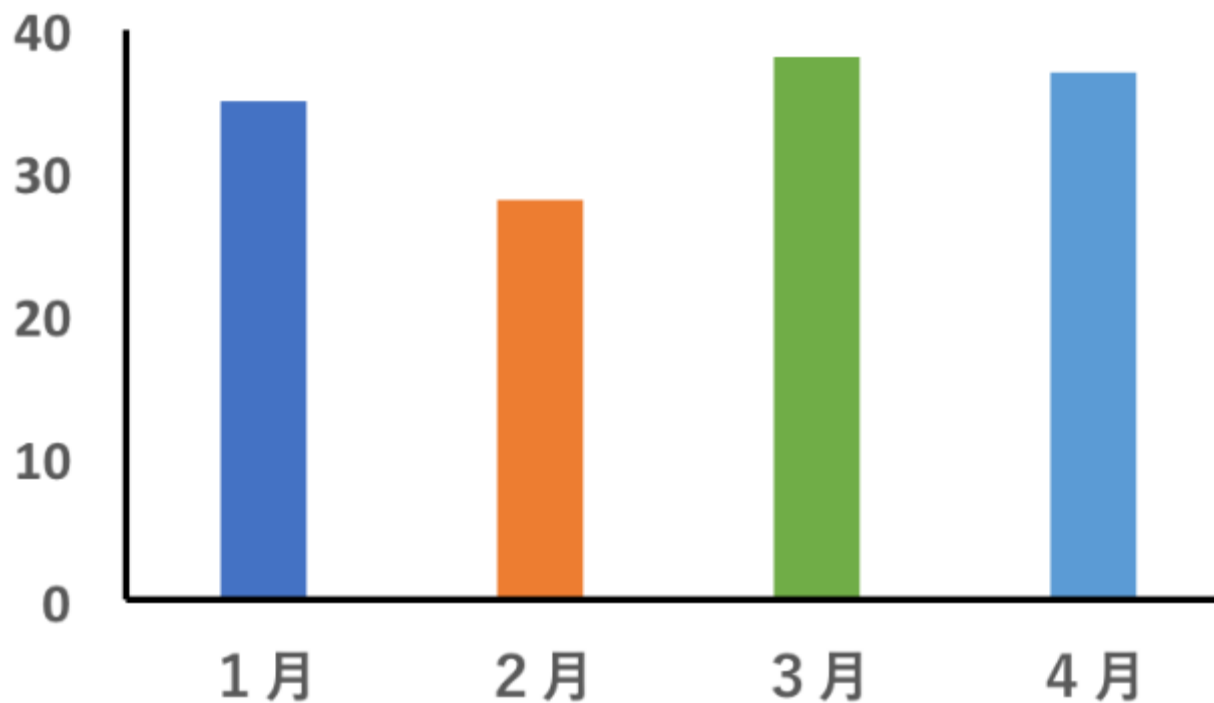
その翌月の売上



キャンペーンが功を奏して、3月の売上は今年一番を記録した！



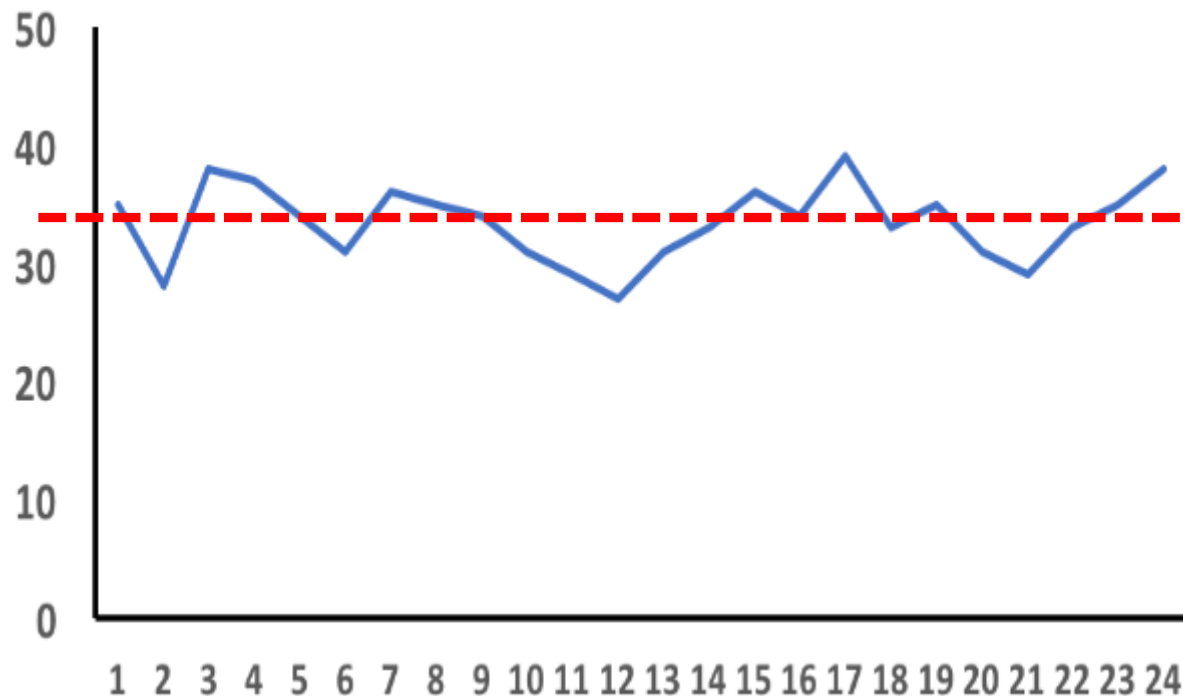
その翌翌月の売上



4月は惜しくも3月よりわずかだけしたまりました。しかし、来月こそは一致団結して、最高売上を狙います。



2年が経過して、振り返ってみると

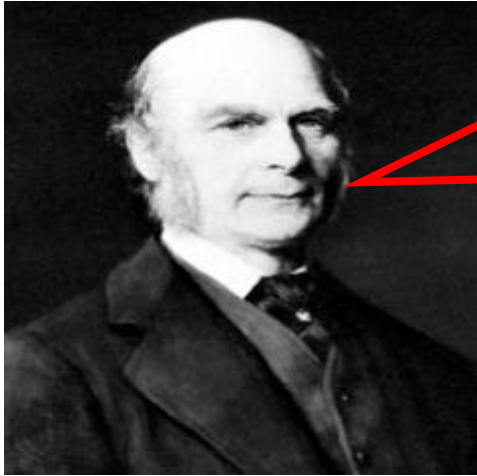


「平均への回帰」

チームで課題を共有し、毎月新しい方針を設定して、対応してるはずなのに？



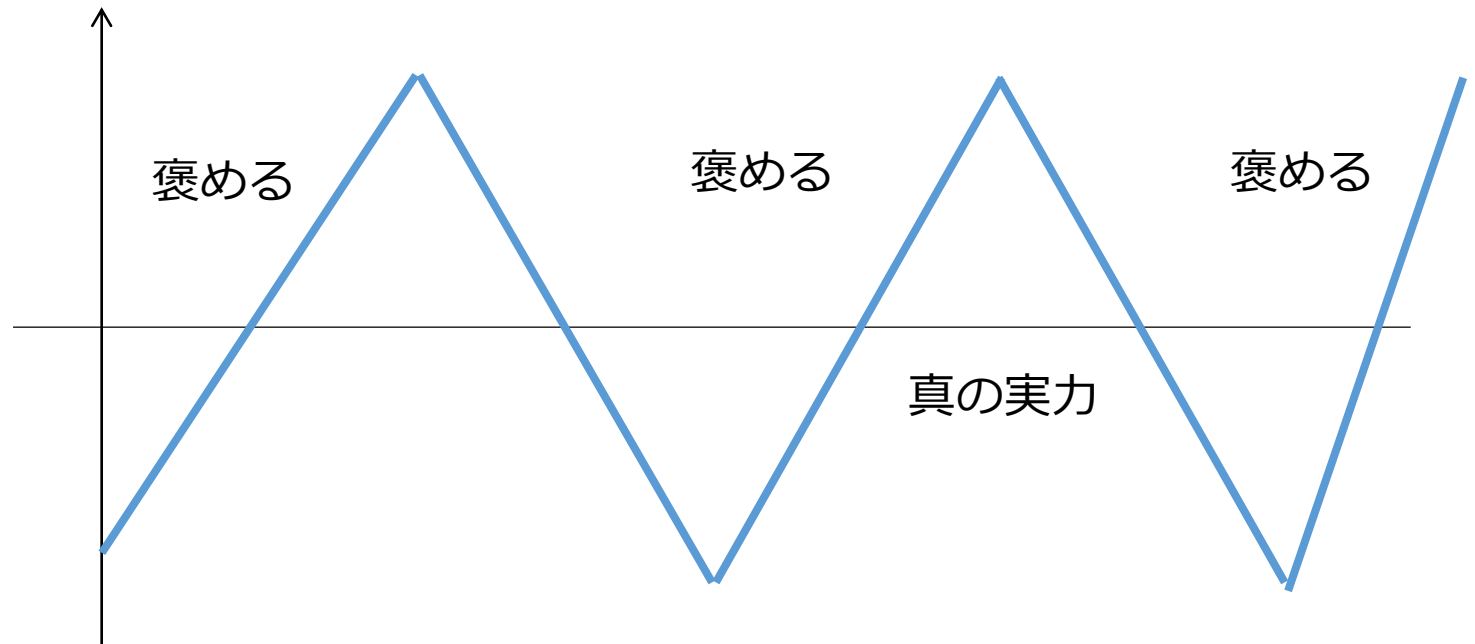
平均への回帰



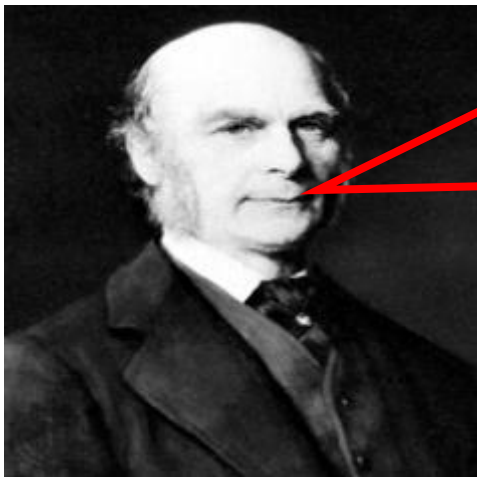
フランシス・ゴルドン
(1822～1911)

世の中の大半の物事はある平均を持っていて、何もしなければ勝手に平均に「戻る」（回帰する）性質がある。

パフォーマンス



ビジネスにおける目標？

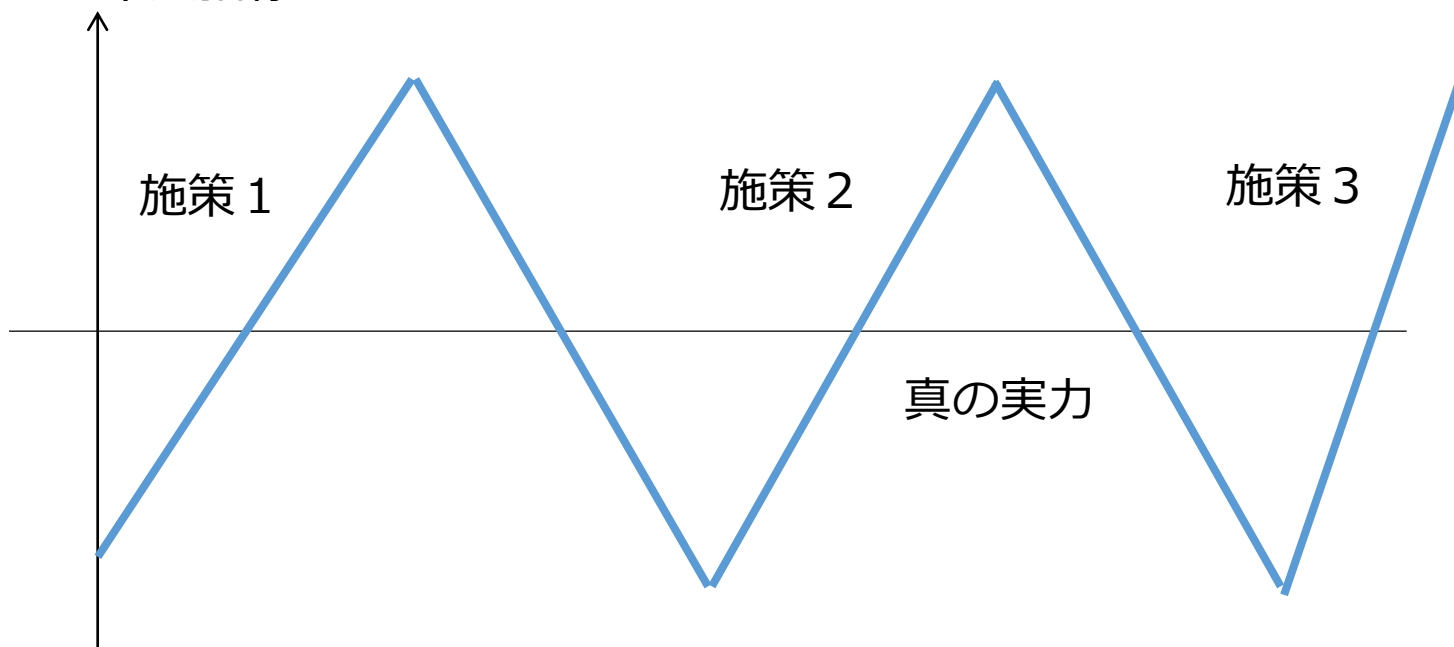


フランシス・ゴルドン
(1822～1911)

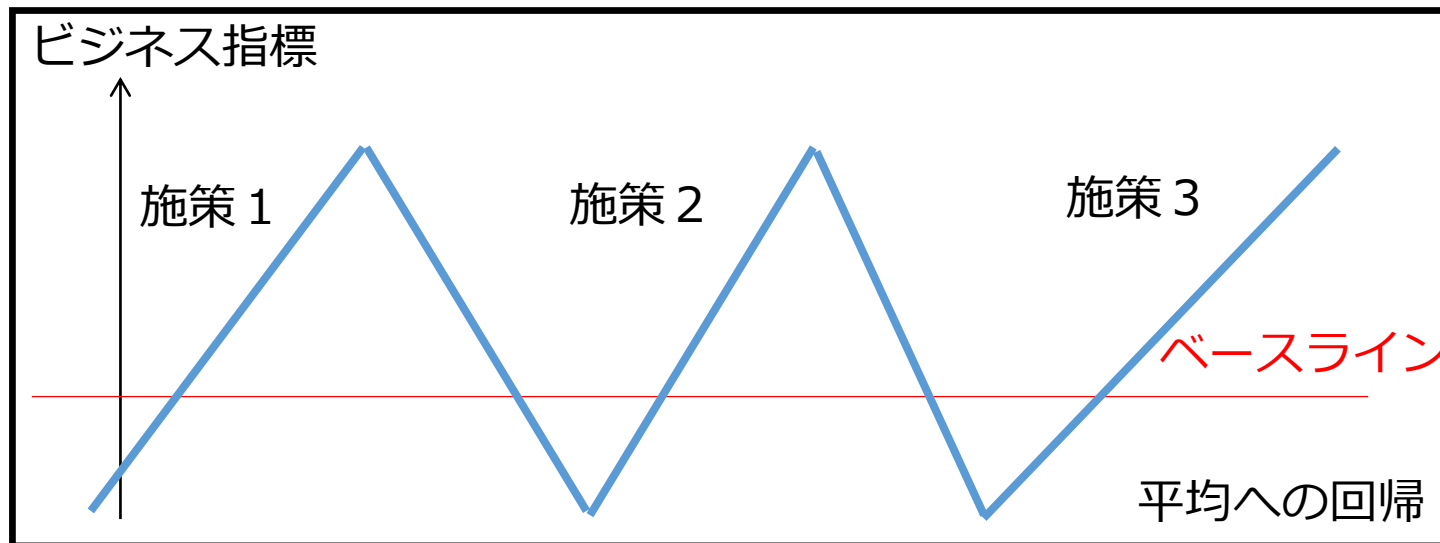
世の中の大半の物事はある平均を持っていて、何もしなければ勝手に平均に「戻る」（回帰する）性質がある。

「平均への回帰」の呪縛から逃れられること

ビジネス指標

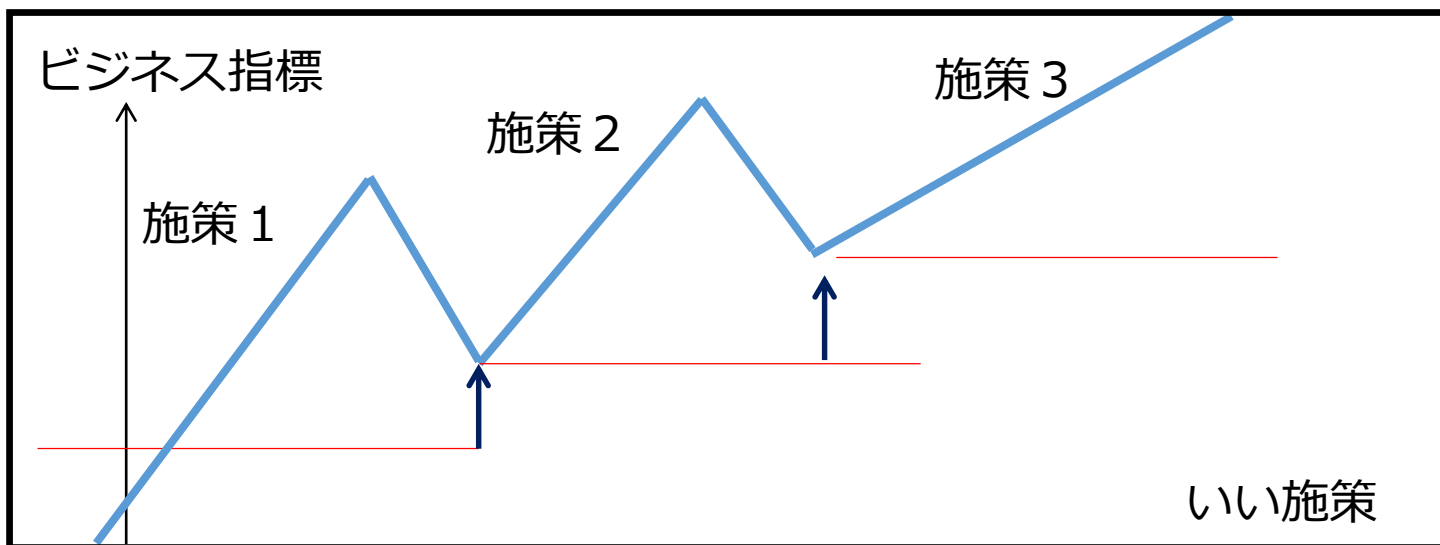


統計学で何ができるのか？

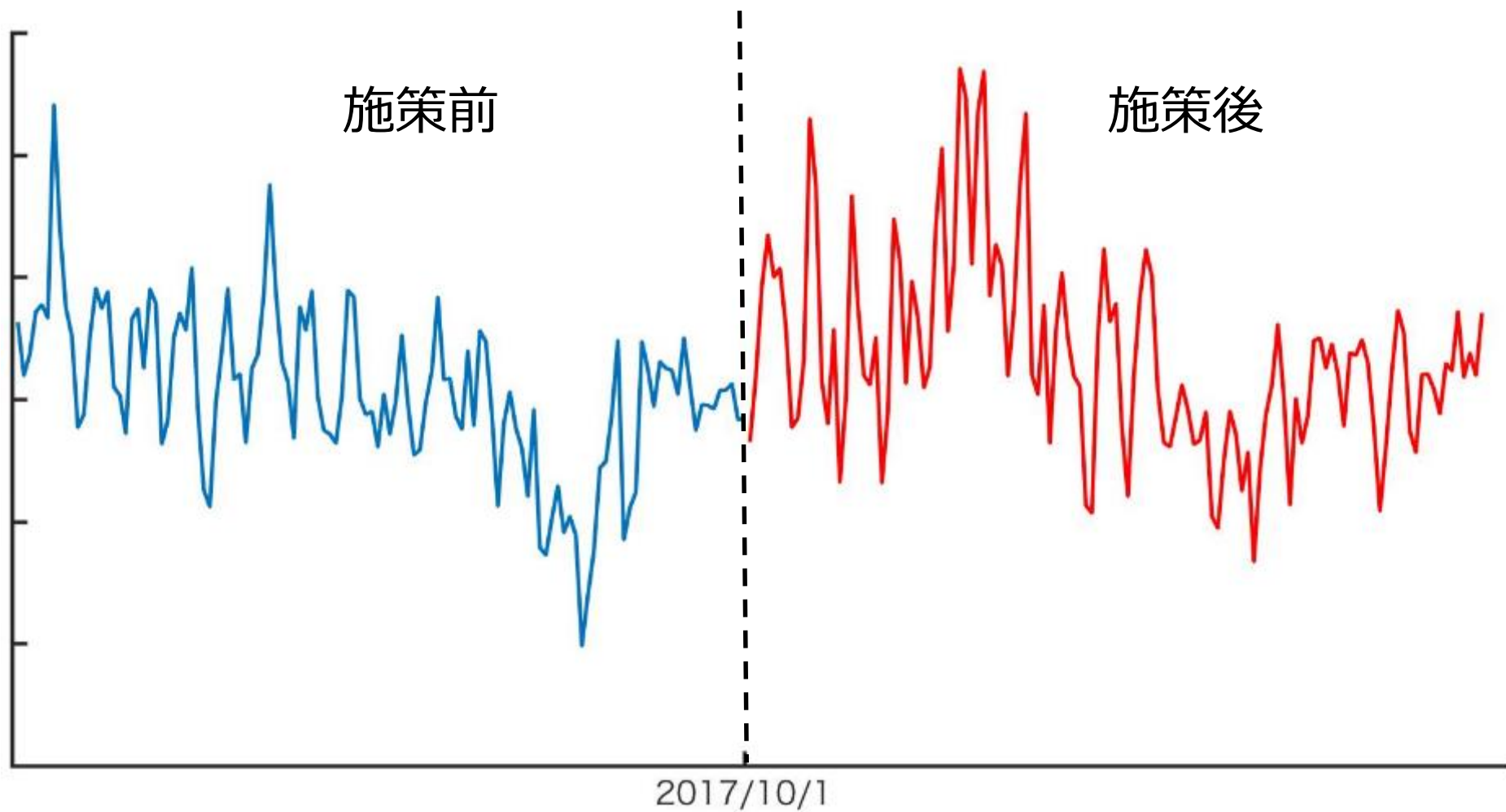


統計学でできること
その1

「施策がうまく
機能しているかを
検証する」



実データ



検定

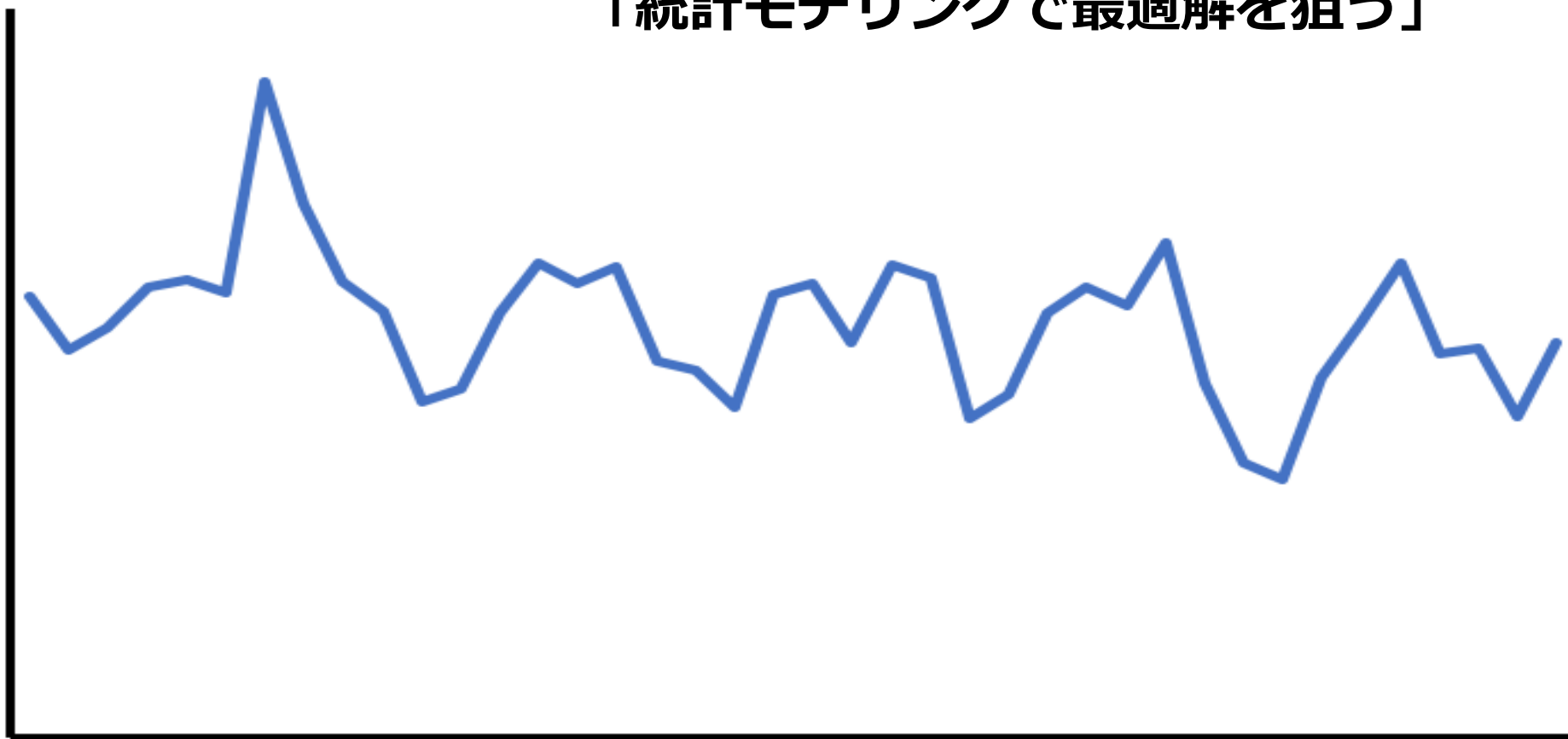
t-検定: 分散が等しくないと仮定した2標本による検定

	売上(施策前)	売上(施策後)
平均	612.6229508	671.9430894
分散	17644.00542	25759.48034
観測数	122	123
仮説平均との差異	0	
自由度	235	
t	-3.152608655	
P(T≤t) 片側	0.000914402	
t 境界値 片側	1.651363544	
P(T≤t) 両側	0.001828803	
t 境界値 両側	1.970110062	

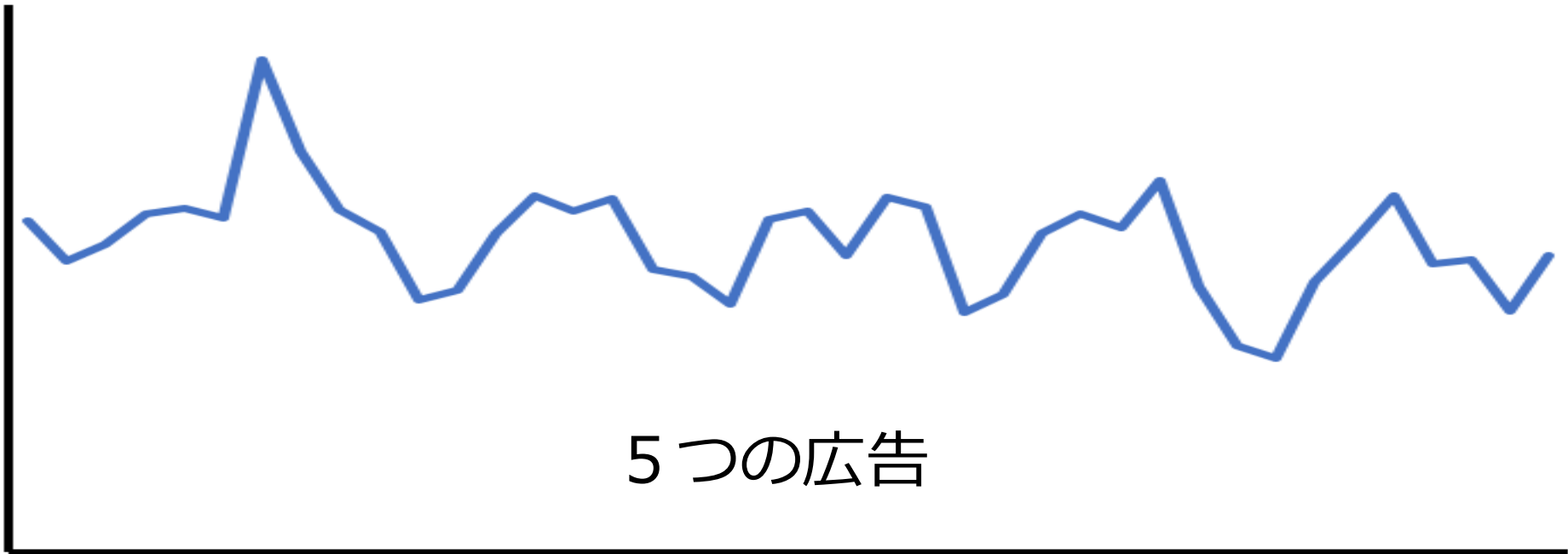
統計学で何ができるのか？

統計学でできること その2

「統計モデリングで最適解を狙う」

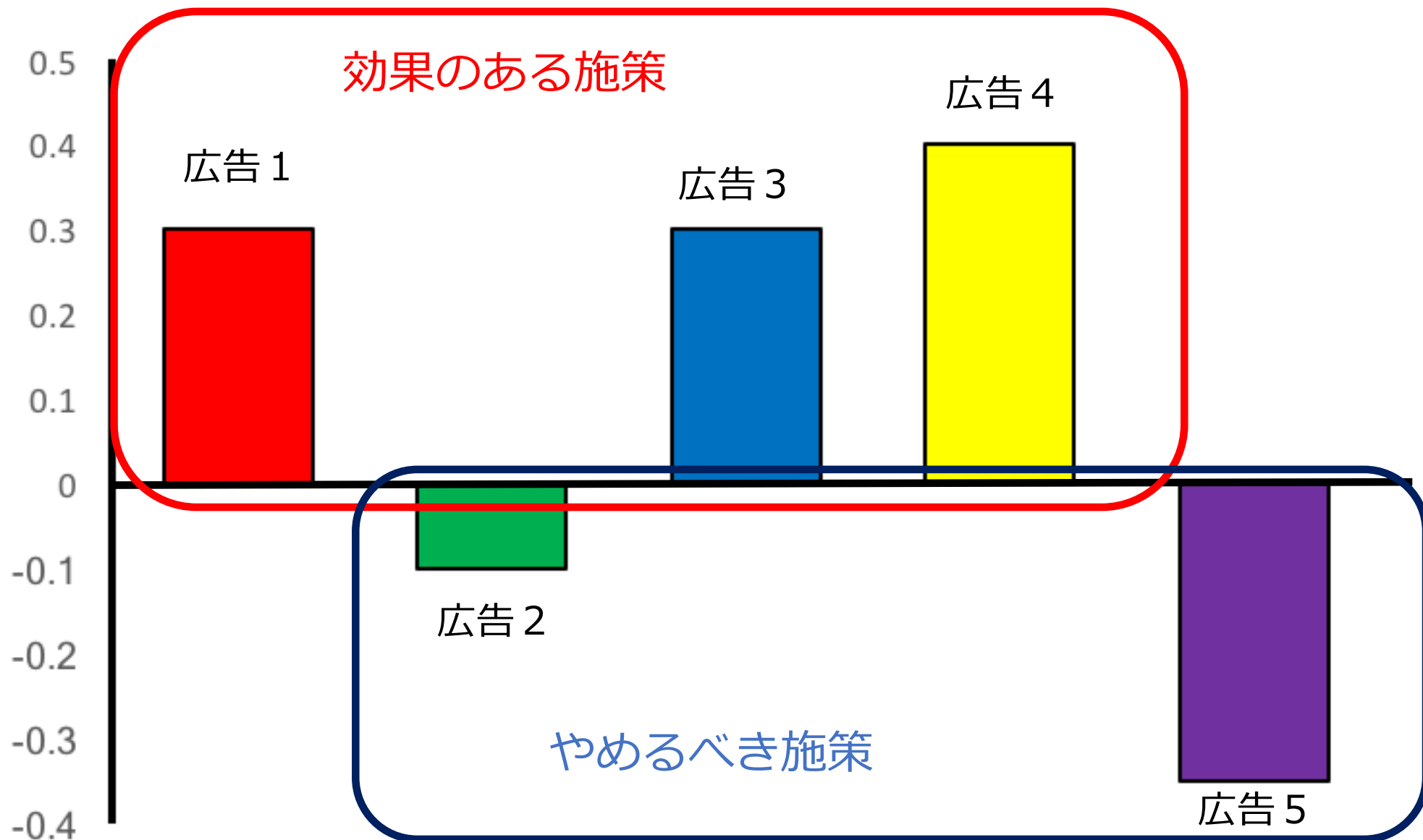


統計モデリングで最適解を狙う

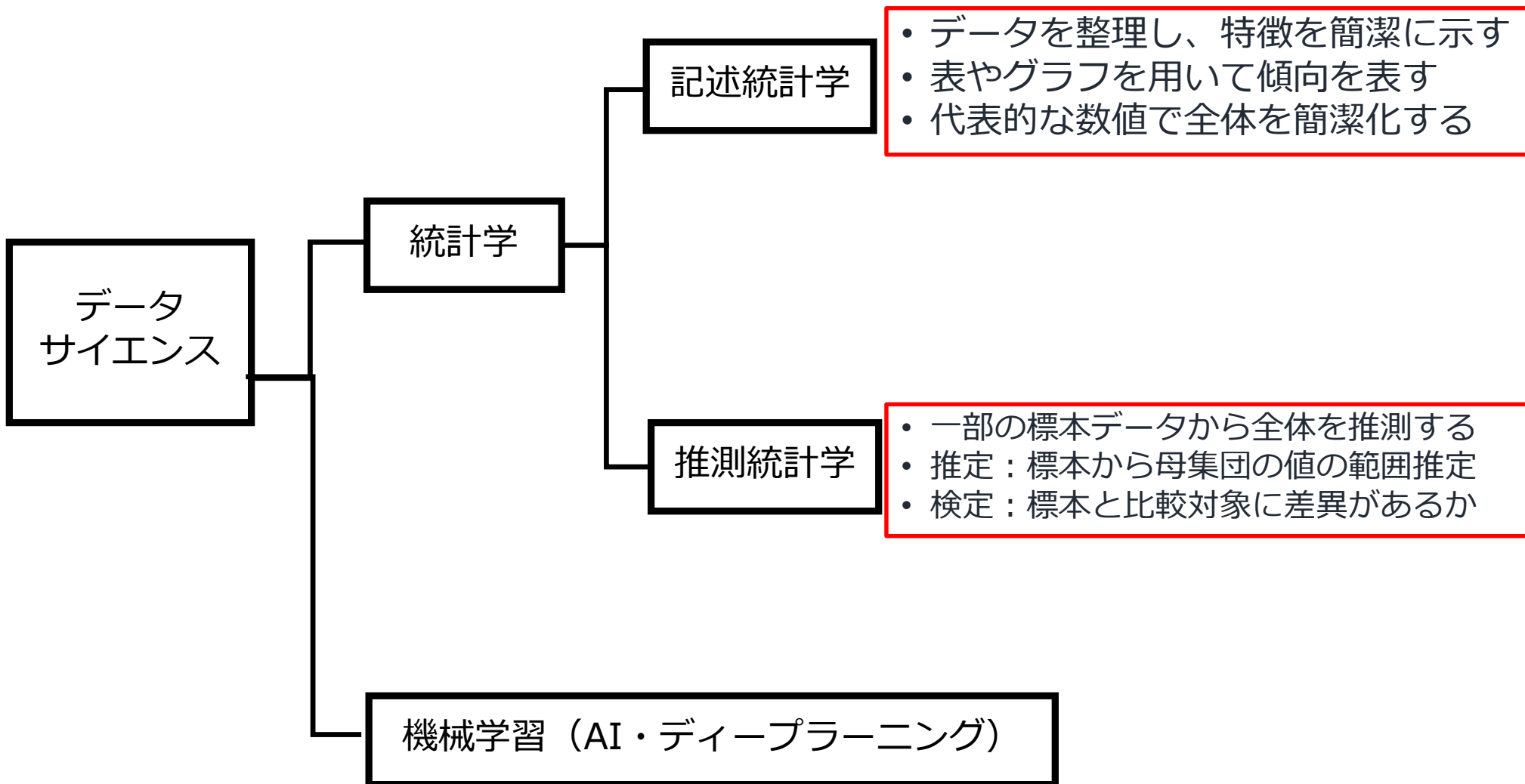


日時	Sales	広告 1	広告 2	広告 3	広告 4	広告 5
6月1日	726	0	15	20	23	0
6月2日	639	23	13	20	12	0
6月3日	674	21	11	20	0	0
6月4日	743	20	12	0	10	12
6月5日	755	21	14	0	1	14
6月6日	733	21	2	0	12	14

施策ごとの影響度の大小関係を見ることができる



データ分析マップ



データ分析の実例

問題：このデータから何がわかるのか？

ID	満足度	他者評価	プロジェクト数	労働時間 (月平均)	労働時間 (会社内)	Work accident	退職・在職	過去5年の 昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium
5637	0.98	0.92	4	175	2	0	在職	無	IT	medium
5305	0.69	0.83	4	264	3	0	在職	無	technical	low
4823	0.66	0.85	3	266	5	0	在職	無	sales	low
9335	0.79	0.49	4	163	3	0	在職	無	sales	high
12400	0.1	0.87	6	250	4	0	退職	無	sales	low
12205	0.87	0.9	5	254	6	0	退職	無	support	low

データの分類

ID	満足度	他者評価	プロジェクト数	労働時間 (月平均)	労働時間 (会社内)	Work accident	退職・在職	過去5年の 昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium

数量データ

- 平均値
- 中央値
- 最大値
- 最小値
- 標準偏差
- 25%、75点
- ヒストグラム

質的データ

- 円グラフ
- クロス集計

虎の巻（データ分析）

データを分析する前に

何を目的として分析するのか？

データを分析するとは

データの要約

データ間の関係性

予測する

結果の検証

問題解決のための哲学

分解と統合

虎の巻（データ分析）

データを分析する前に

何を目的として分析するのか？

ID	満足度	他者評価	プロジェクト数	労働時間 (月平均)	労働時間 (会社内)	Work accident	退職・在職	過去5年の 昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium

社員は会社にいるのだろうか？

（レベル1 集計）

虎の巻（データ分析）

量的データの集計

ID	満足度	他者評価	プロジェクト数	労働時間 (月平均)	労働時間 (会社内)	Work accident	退職・在職	過去5年の 昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium

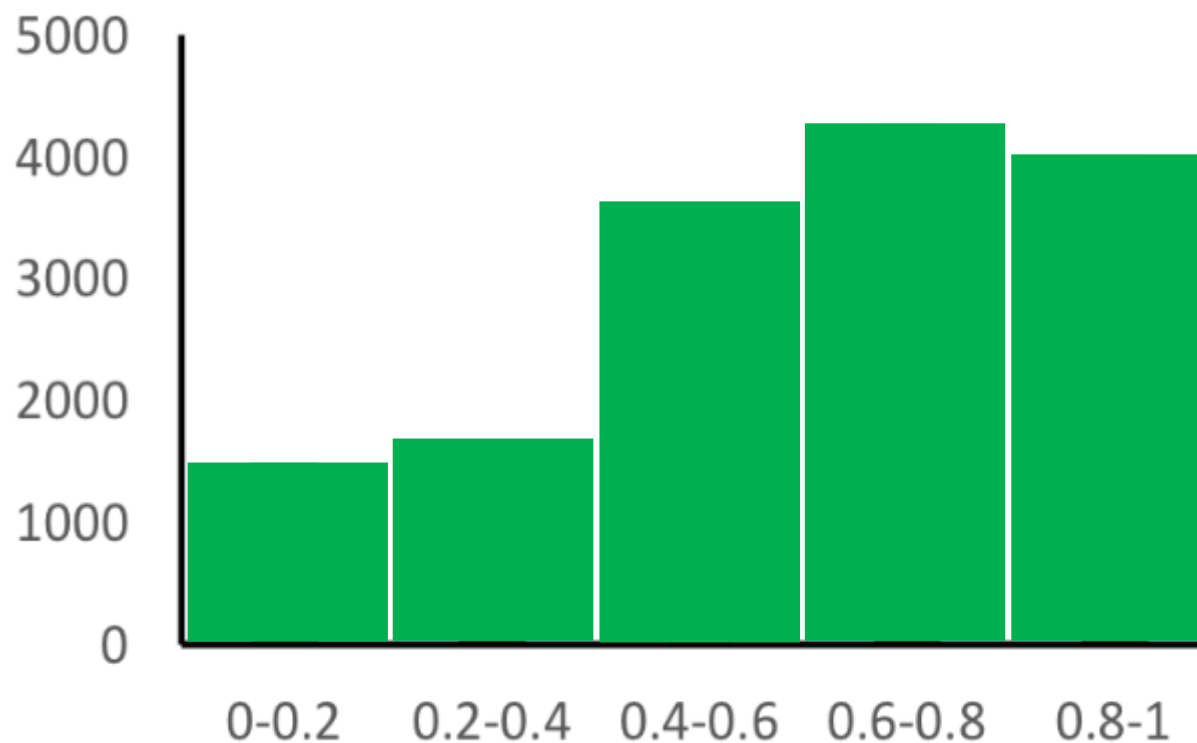
虎の巻（データ分析）

データを分析するとは

データの要約

満足度

データ区間	頻度
0～0.2	1478
0.2～0.4	1646
0.4～0.6	3605
0.6～0.8	4268
0.8～1.0	4002



虎の巻（データ分析）

データを分析する前に

何を目的として分析するのか？

ID	満足度	他者評価	プロジェクト数	労働時間 (月平均)	労働時間 (会社内)	Work accident	退職・在職	過去5年の 昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium

このデータからどの社員が退職するか予測することは可能なのか？

(レベル2 検定)

(レベル3 予測モデルの設計)

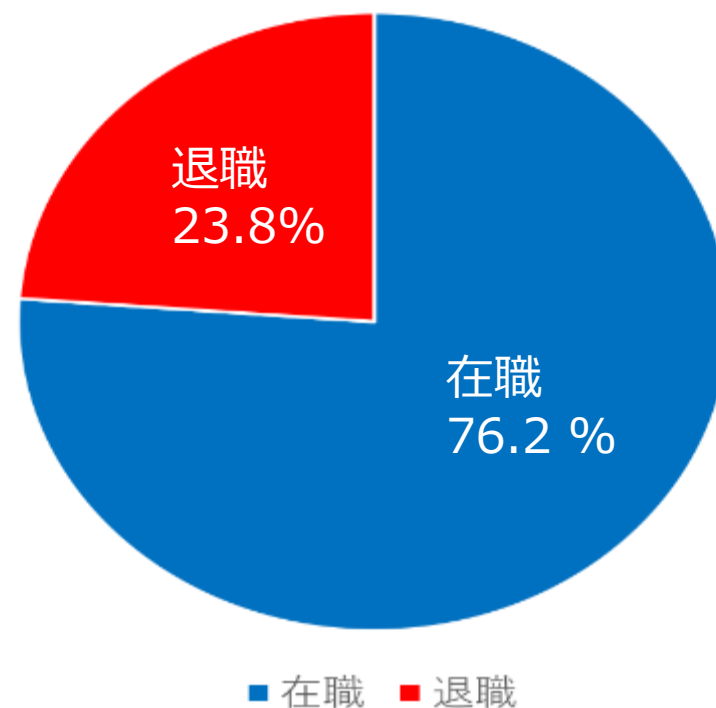
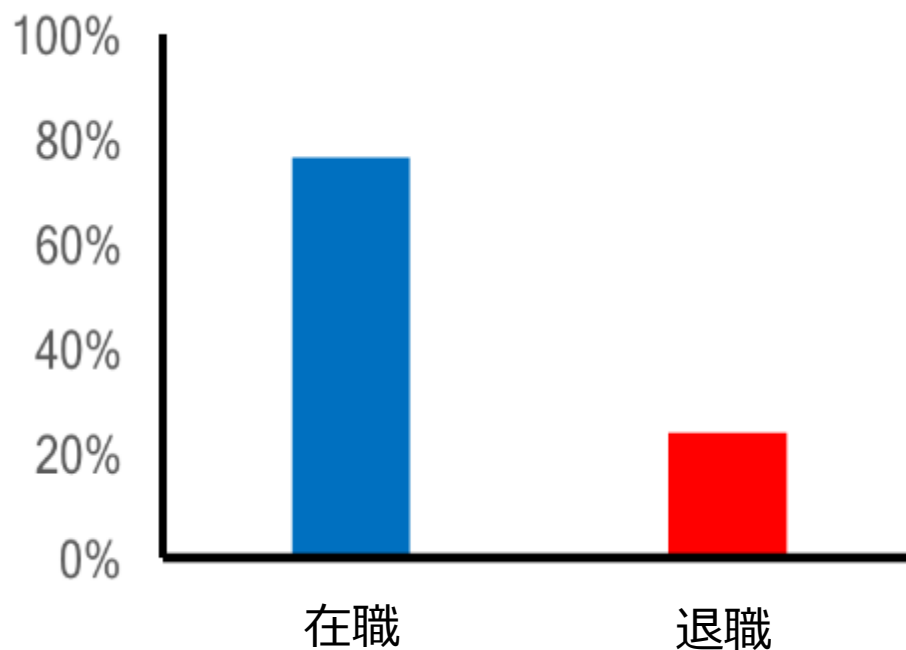
虎の巻（データ分析）

質的データの集計

ID	満足度	他者評価	プロジェクト数	労働時間 (月平均)	労働時間 (会社内)	Work accident	退職・在職	過去5年の 昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium

データの可視化

退職	在職
3571	11428
23.8%	76.2%



虎の巻（データ分析）

データを分析するとは

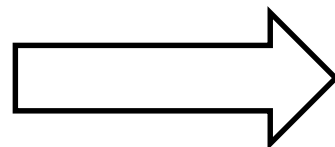
データ間の関係性

ID	満足度	他者評価	プロジェクト数	労働時間 (月平均)	労働時間 (会社内)	Work accident	退職・在職	過去5年の 昇進	所属部署	給料
1019	0.36	0.47	2	136	3	0	退職	無	accounting	low
6830	0.68	0.51	5	158	3	0	在職	無	technical	medium
9653	0.53	0.64	2	109	3	0	在職	無	hr	medium
12208	0.78	0.87	4	228	5	0	退職	無	support	low
4816	0.92	0.56	4	170	3	0	在職	無	marketing	medium

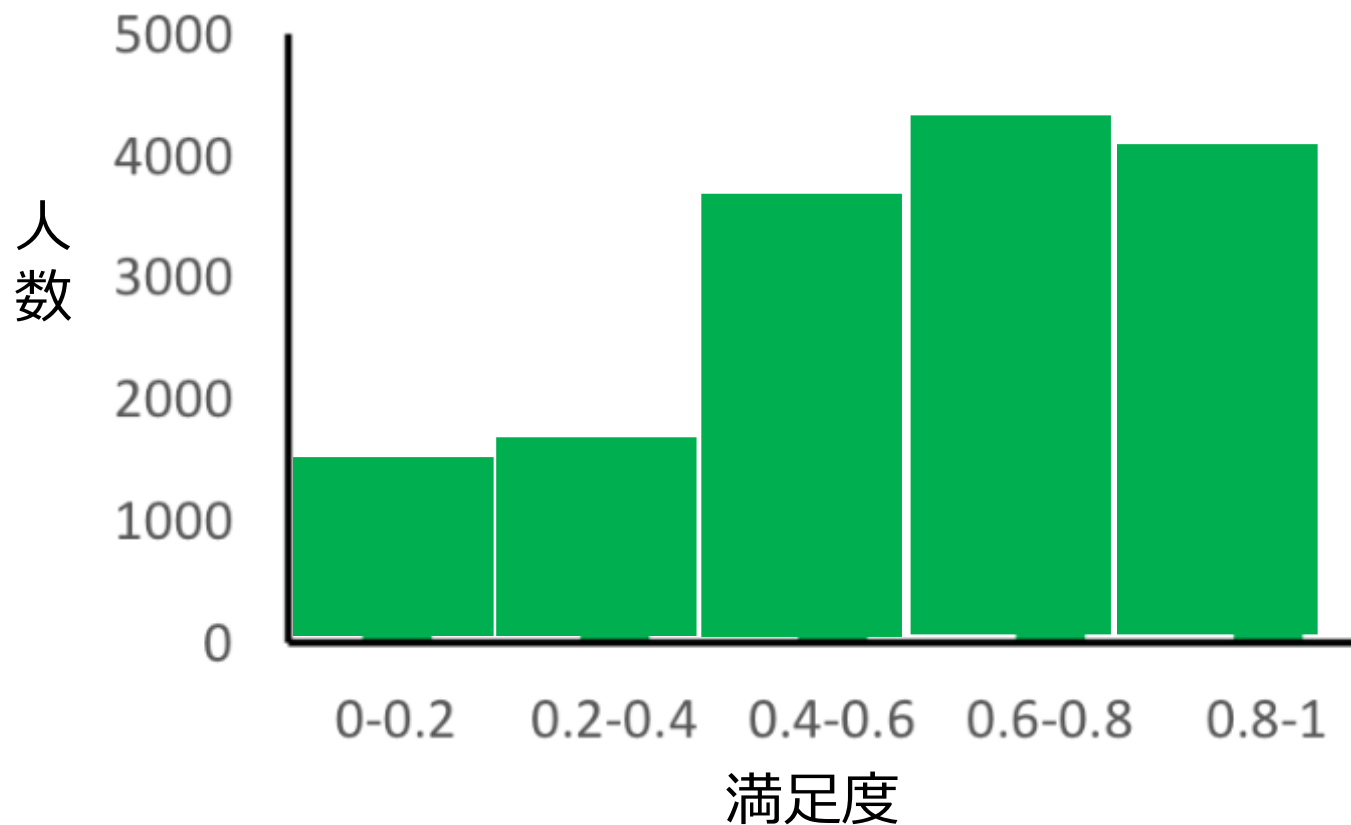
虎の巻（データ分析）

問題解決のための哲学

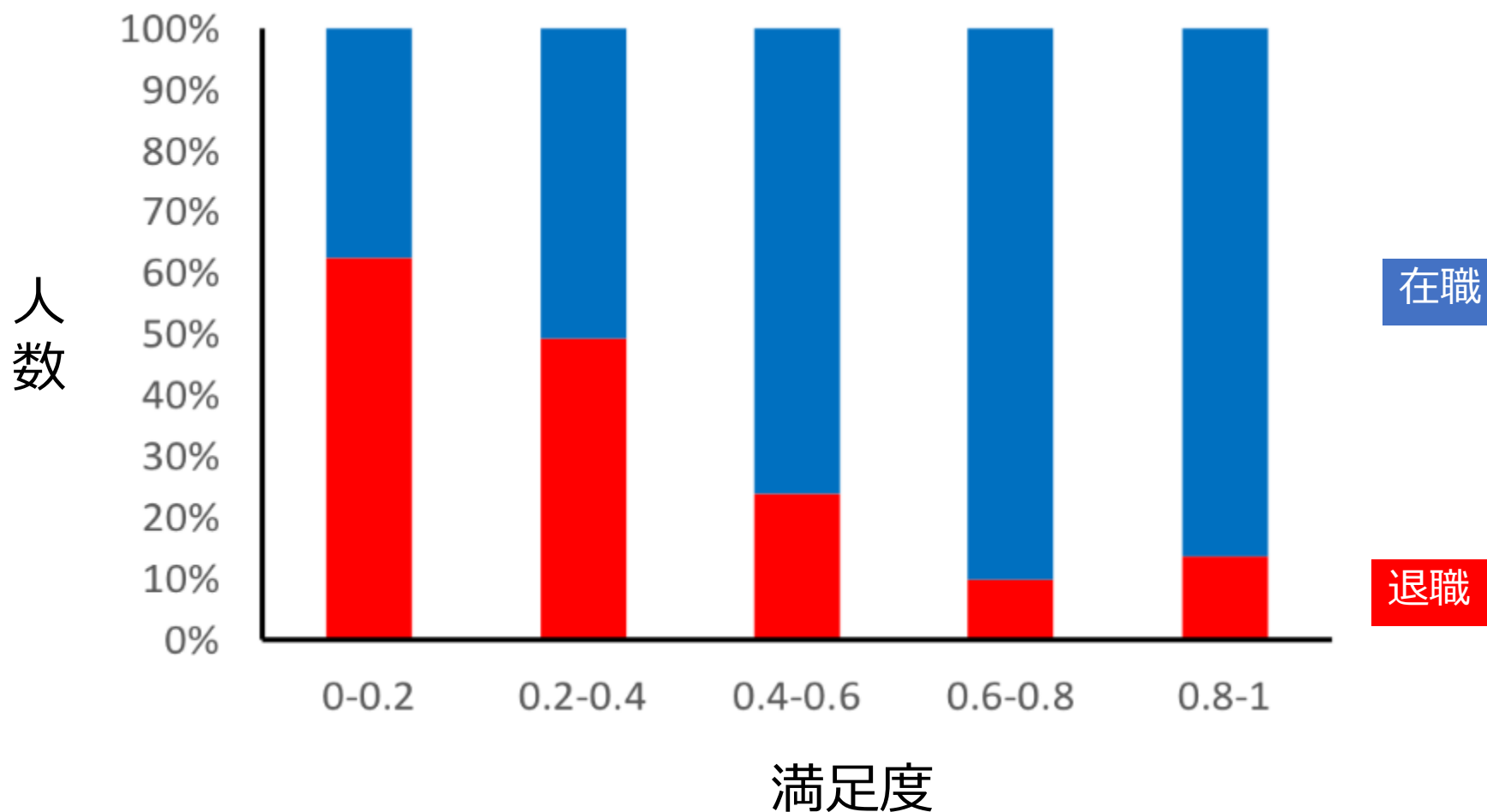
分解と統合



具体的にどうする？



虎の巻（データ分析）



データの可視化(バブルチャート)

3つの変数の関係(満足度、部署、退職率)?

満足度	他者評価	プロジェクト数	労働時間 (月平均)	労働時間 (会社内)	Work accident	退職か在职	過去5年 昇進(有無)	所属部署	給料
0.58	0.55	4	202	3	0	在职	無	IT	medium
0.67	0.74	3	226	3	0	在职	無	product_mng	low
0.11	0.91	7	287	4	0	退職	無	sales	low
0.37	0.5	2	135	3	0	退職	無	product_mng	low
0.93	0.79	5	241	4	0	在职	無	marketing	high
0.4	0.38	3	280	2	0	在职	無	marketing	low
0.23	0.64	5	150	5	0	在职	無	hr	medium
0.83	0.98	5	189	4	1	在职	無	management	low
0.2	0.58	3	209	5	0	在职	無	hr	medium
0.95	0.7	4	267	3	1	在职	無	technical	low
0.11	0.8	6	282	4	0	退職	無	technical	medium
0.7	0.5	6	214	5	0	在职	無	support	medium
0.43	0.51	5	168	4	0	在职	無	product_mng	medium
0.46	0.75	6	276	6	0	在职	無	support	low
0.67	0.8	4	137	2	0	在职	無	support	medium
0.63	0.88	4	260	2	0	在职	無	sales	low
0.99	0.92	5	213	2	0	在职	無	hr	high
0.24	0.94	4	146	4	0	在职	無	product_mng	medium
0.55	0.82	4	134	6	0	在职	無	technical	medium

データの可視化(バブルチャート)

3つの変数の関係(満足度、部署、退職率)？

	accounting	hr	IT	management	marketing	product_mng	sales	support	technical
0.8-1									
0.6-0.8									
0.4-0.6									
0.2-0.4									
0-0.2									

データの可視化(バブルチャート)

3つの変数の関係(満足度、部署、退職率)？

	accounting	hr	IT	management	marketing	product_mng	sales	support	technical
0.8-1	30	28	43	14	33	38	150	91	129
0.6-0.8	12	22	26	11	20	30	128	72	62
0.4-0.6	59	80	66	25	73	56	311	154	177
0.2-0.4	44	35	50	13	36	34	186	103	126
0-0.2	59	50	88	28	41	40	239	135	203

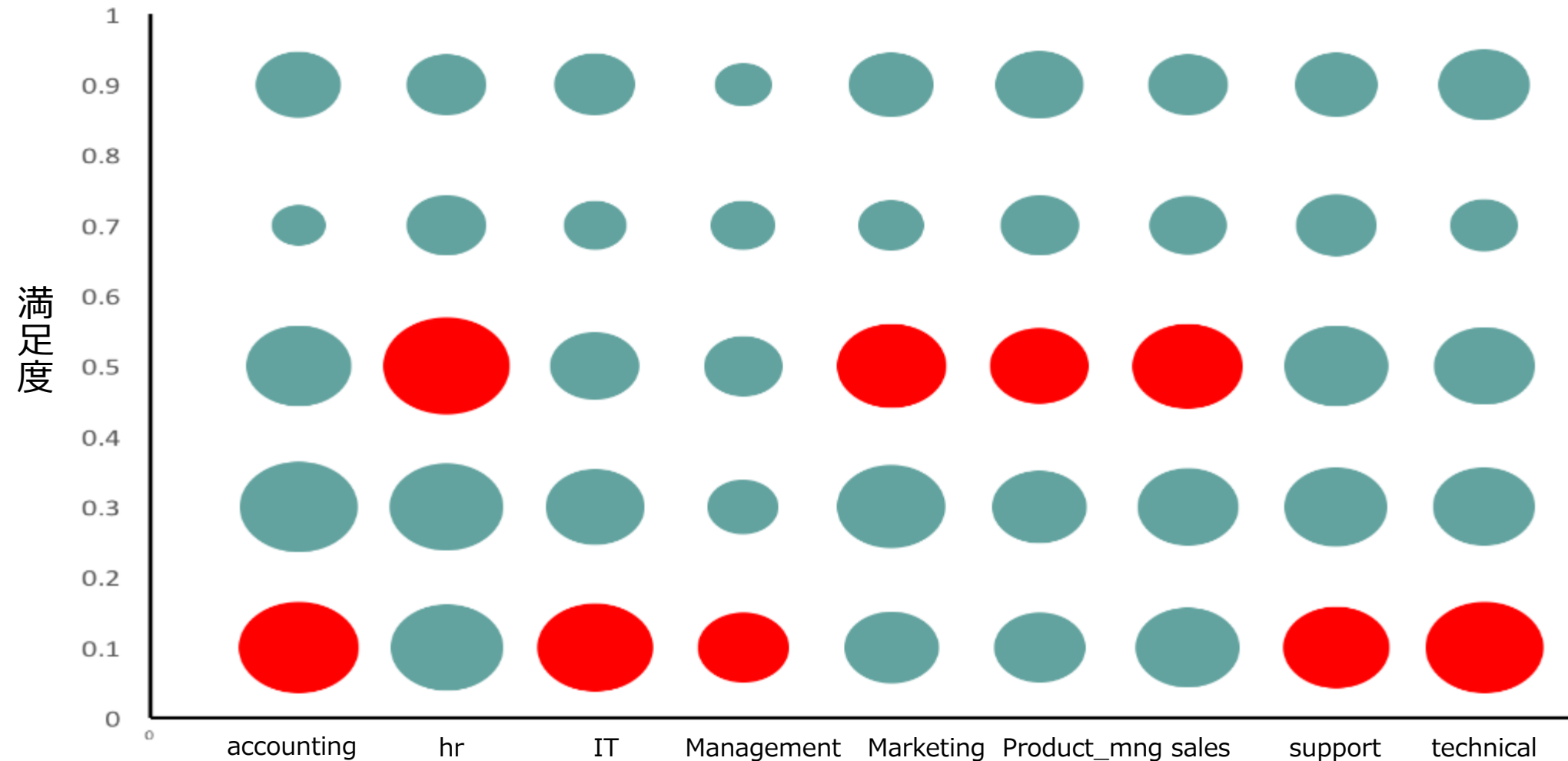
データの可視化(バブルチャート)

3つの変数の関係(満足度、部署、退職率)？

	accounting	hr	IT	management	marketing	product_mng	sales	support	technical
0.8-1	30	28	43	14	33	38	150	91	129
0.6-0.8	12	22	26	11	20	30	128	72	62
0.4-0.6	59	80	66	25	73	56	311	154	177
0.2-0.4	44	35	50	13	36	34	186	103	126
0-0.2	59	50	88	28	41	40	239	135	203

データの可視化(バブルチャート)

3つの変数の関係(満足度、部署、退職率)？



エクセルハンズオン

- 基本統計量の計算
- 移動平均法
- データの正規化
- クロス集計

平均と標準偏差

- 平均と標準偏差について理解を深める
- 平均と標準偏差を使った分析
- 平均を使った分析事例（移動平均法）

どっちが優秀なのか？

問題

テストの結果から優秀な人材を選別しなければならない時、あなたはどのように選ぶでしょうか？

- 1 点数をそのまま比べる
- 2 点数と平均点との差を比べる
- 3 標準偏差（偏差値）を比べる

どっちが優秀なのか？

第4回 不合判定テスト成績表

	第1回 不合判定テスト (15/9/21)				第2回 不合判定テスト (15/10/19)			
教科	得点	平均点	偏差値	順位/受験者	得点	平均点	偏差値	順位/受験者
算数	115	83.9	61	793 / 6763	115	79.3	64	476 / 7507
国語	66	73.4	46	4300 / 6763	129	86.2	69	138 / 7507
社会	66	53.0	58	1150 / 5617	66	52.6	59	1159 / 6134
理科	55	44.4	57	1282 / 5664	72	51.1	65	454 / 6191
2教科	181	157.2	56	2089 / 6763	244	165.5	68	198 / 7507
3教科	236	208.9	55	1746 / 5664	316	224.4	68	217 / 6191
4教科	302	262.0	56	1556 / 5617	382	277.3	66	315 / 6134

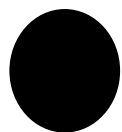
標準偏差

どっちが優秀なのか？

大谷



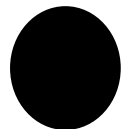
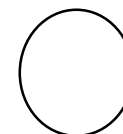
ベーブルース



60点

得点をそのまま比べる

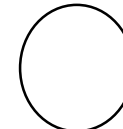
70点



50点

点数と平均点を比べる

50点



どっちが優秀なのか？

大谷



ベーブ・ルース



カール・ピアソン

遺伝の研究をしていました

標準偏差を使って、
集団のばらつきを数字で表す



標準偏差

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}$$

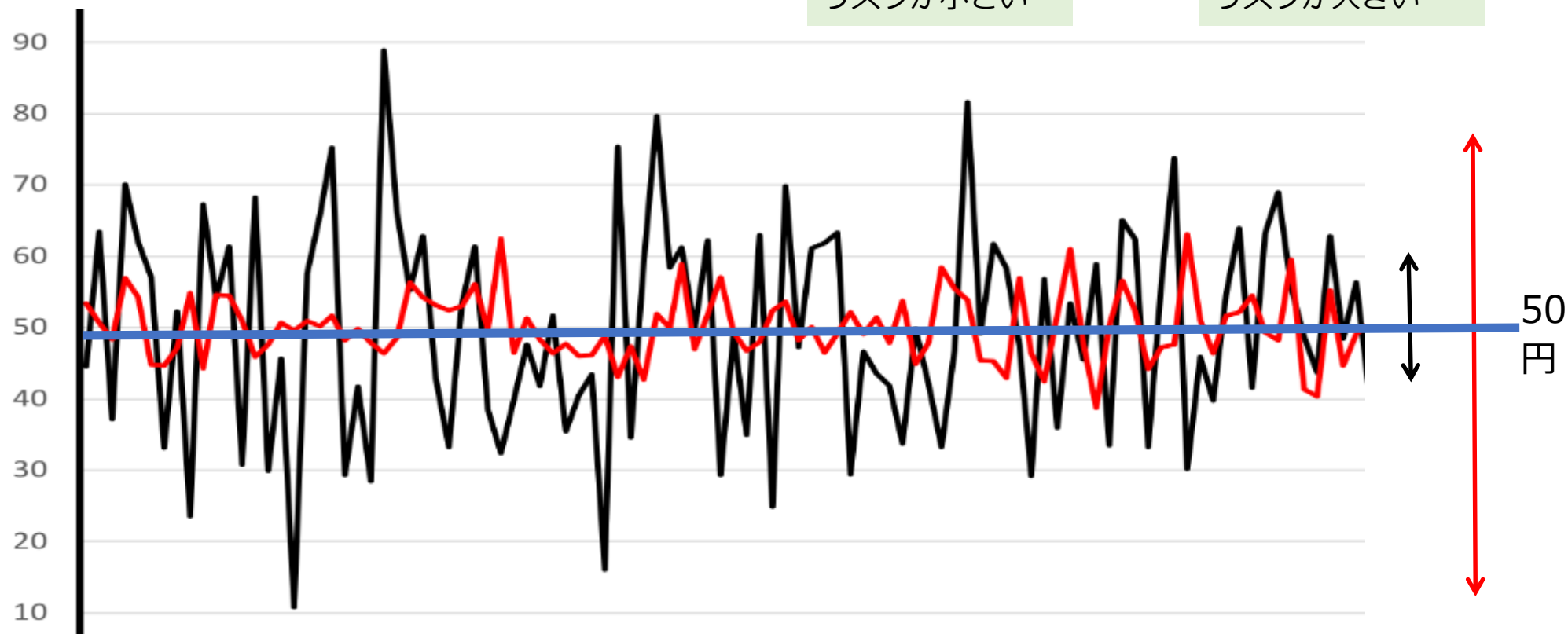
難しく見えるけど、考え方は単純

標準偏差をイメージする

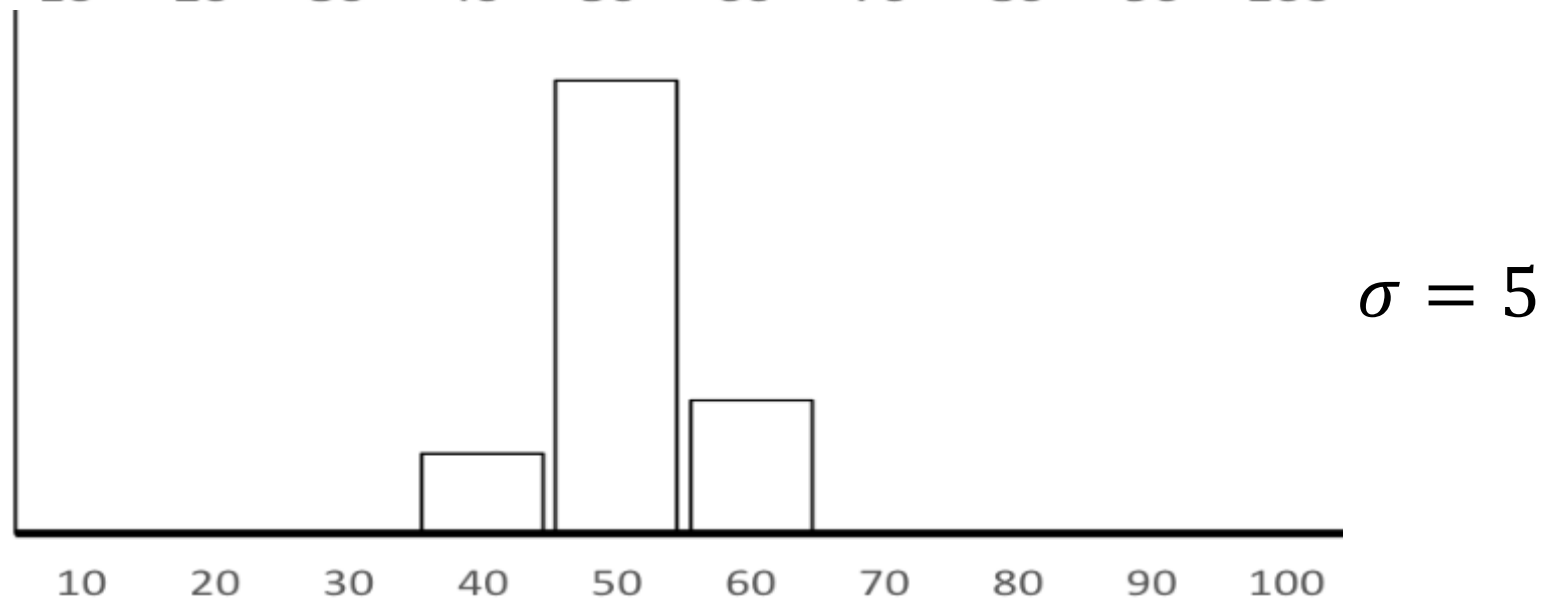
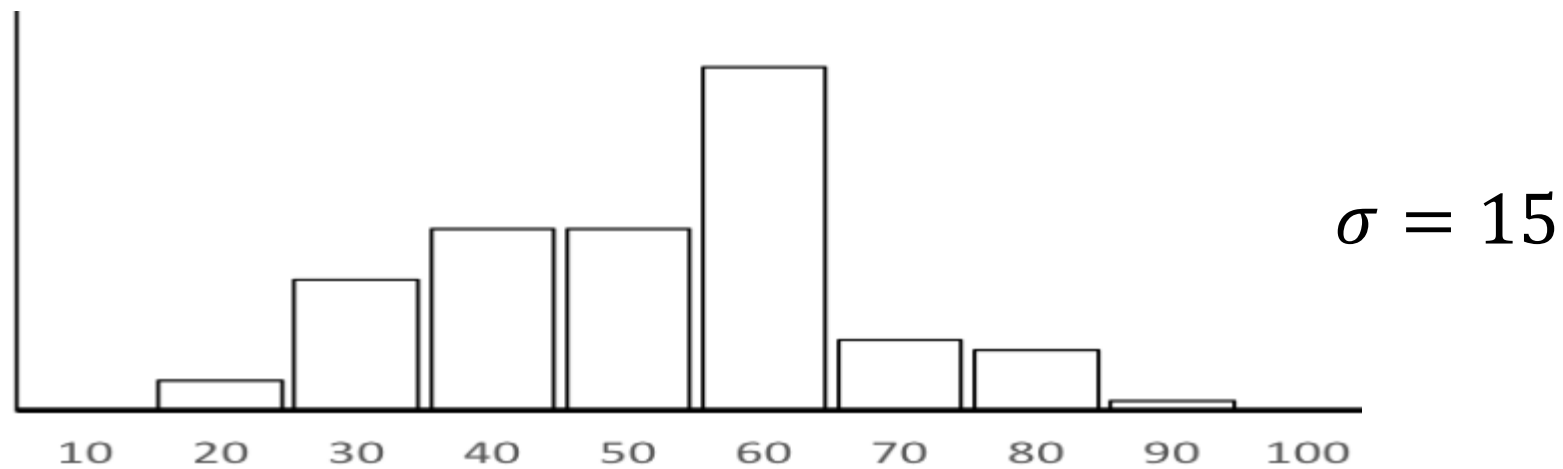
株価Aのばらつき < 株価Bのばらつき

標準偏差が小さい
リスクが小さい

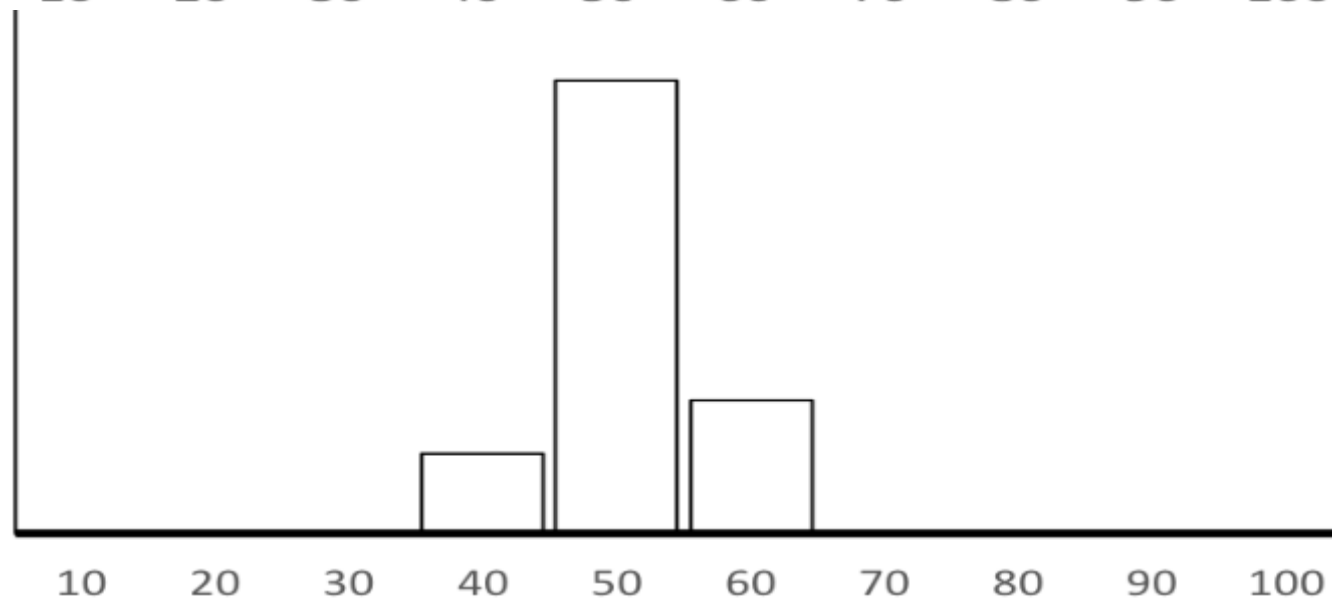
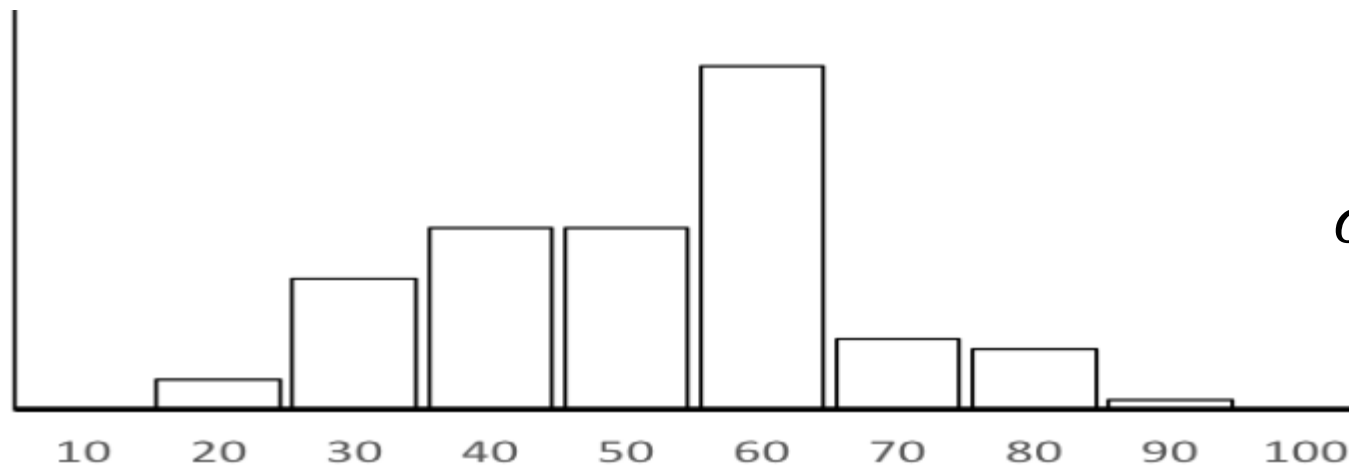
標準偏差が大きい
リスクが大きい



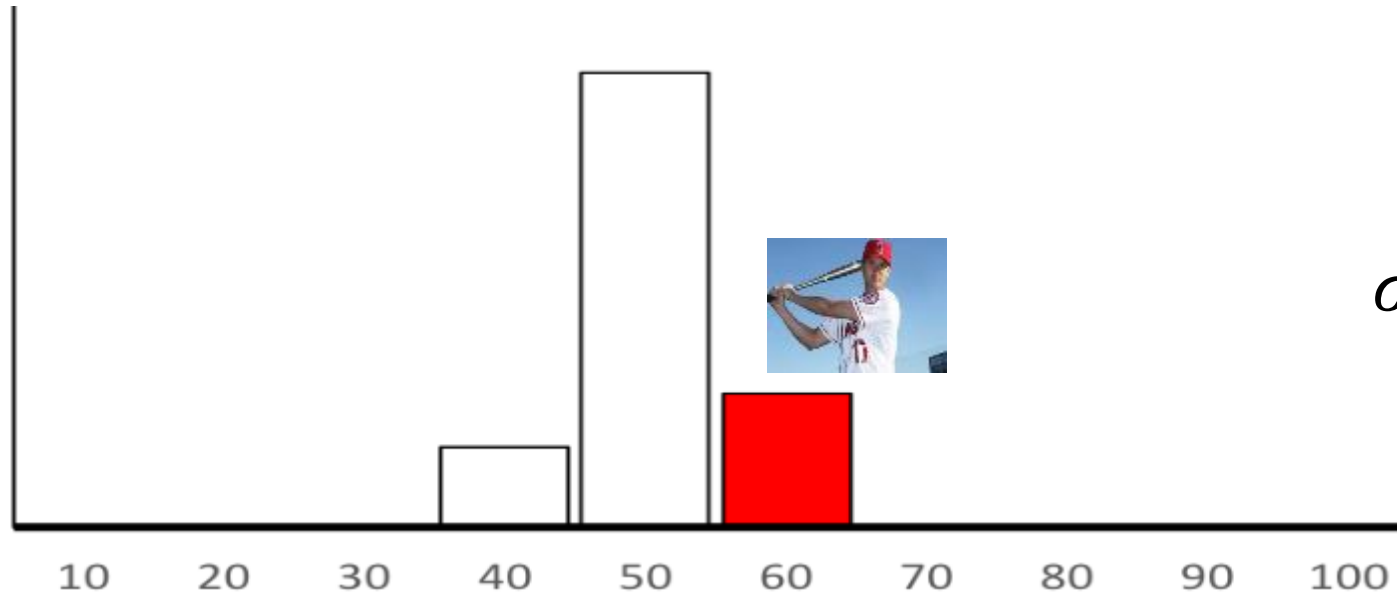
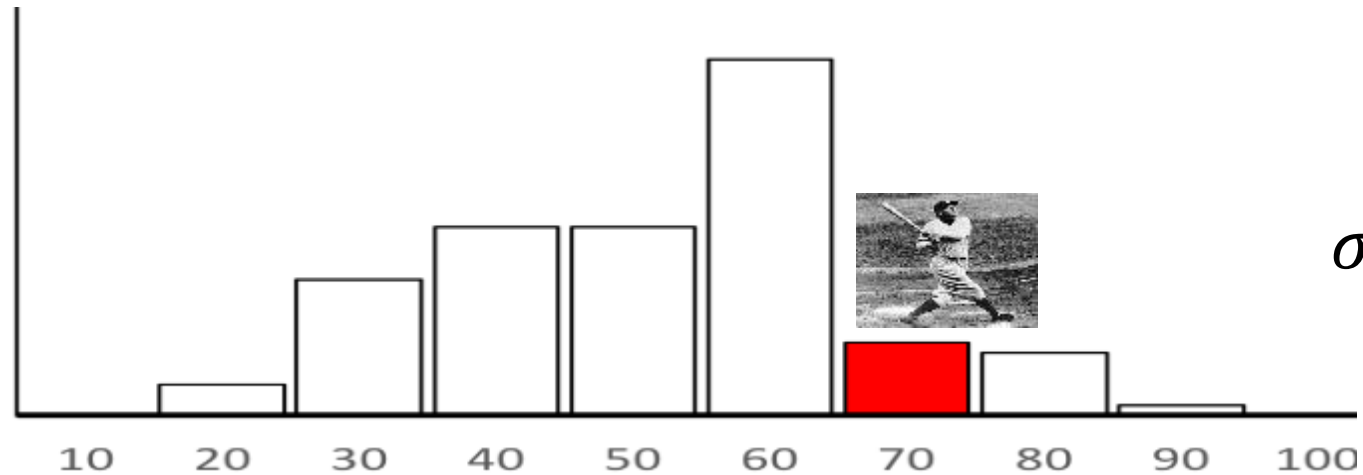
株価のヒストグラム



100人のテスト結果のヒストグラム



100人のテスト結果のヒストグラム

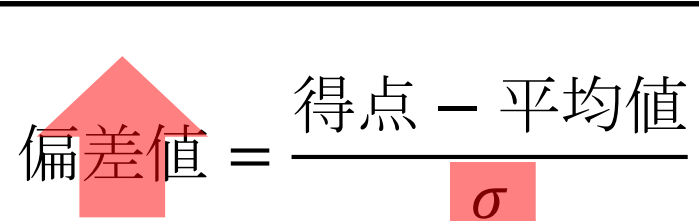


偏差値

$$\text{偏差値} = \frac{\text{得点} - \text{平均値}}{\sigma}$$

偏差値が高くなる条件


$$\text{偏差値} = \frac{\text{得点} - \text{平均値}}{\sigma}$$


$$\text{偏差値} = \frac{\text{得点} - \text{平均値}}{\sigma}$$

偏差値

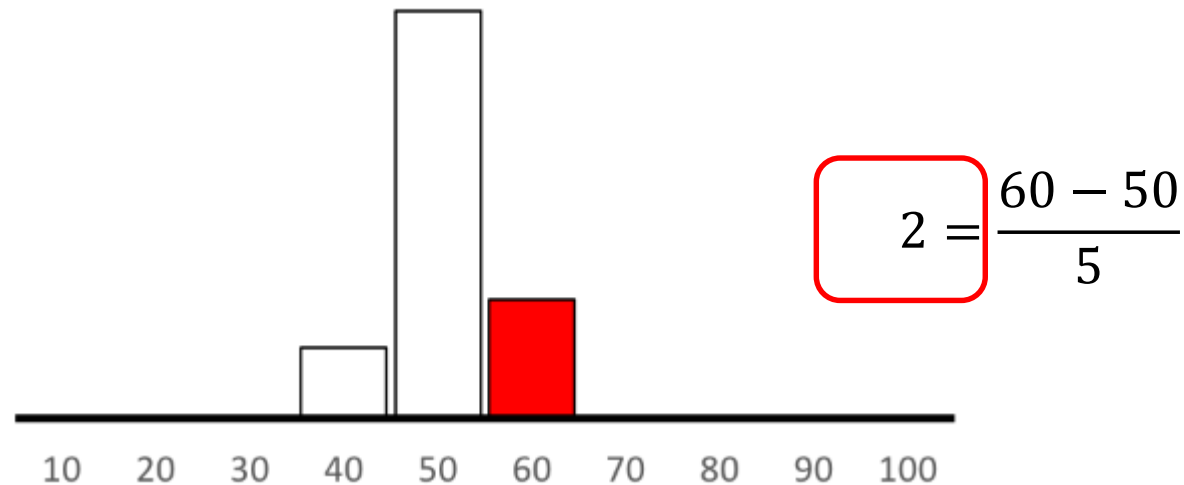
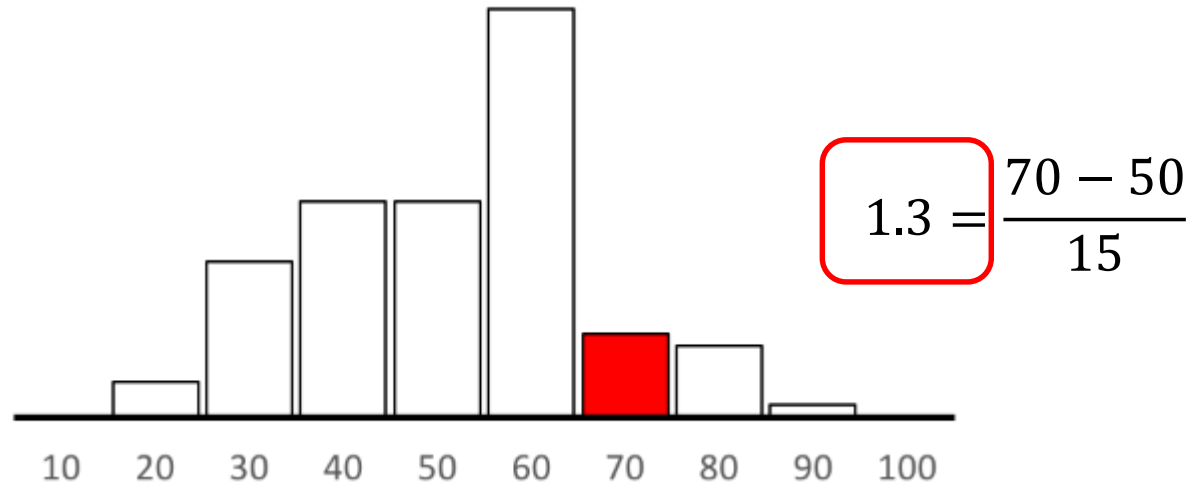


$$\text{偏差値} = \frac{\text{得点} - \text{平均値}}{\sigma} \times 10 + 50$$

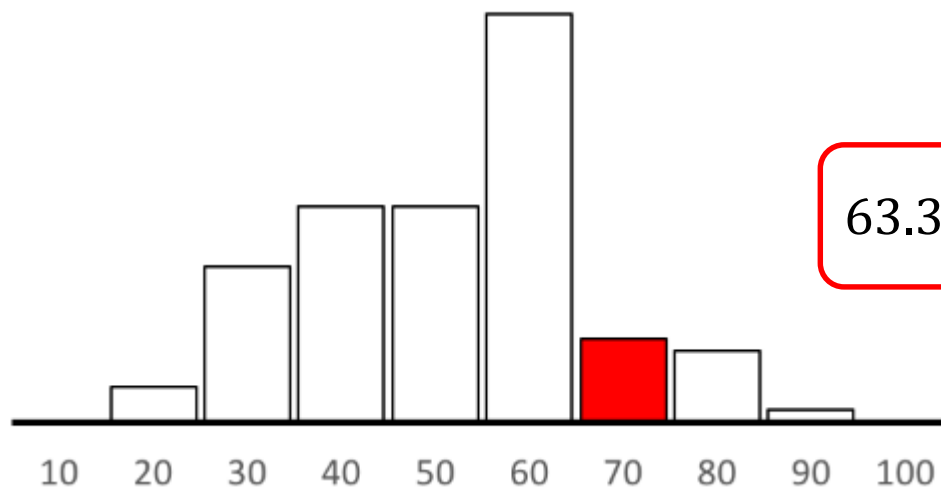


こんな使われ方をするとは。。。

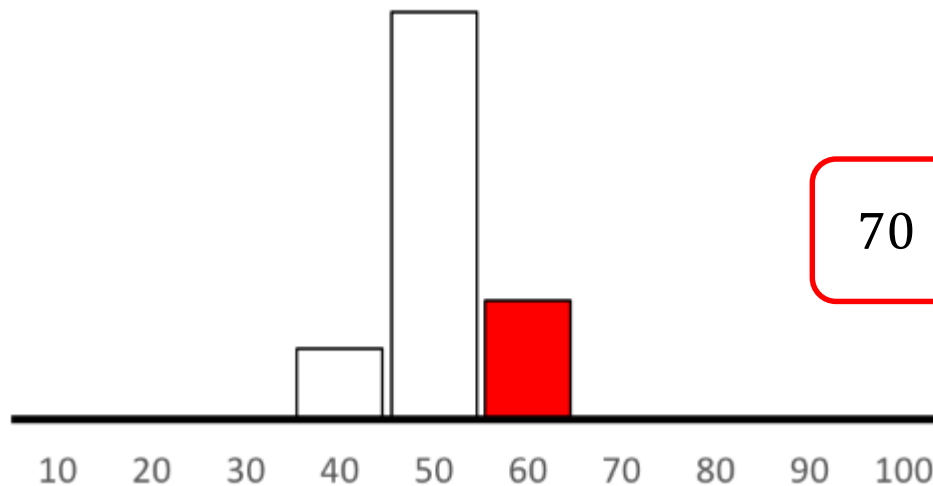
偏差値を求める



偏差値を求める



$$63.3 = \frac{70 - 50}{15} \times 10 + 50$$



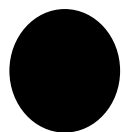
$$70 = \frac{60 - 50}{5} \times 10 + 50$$

どっちが優秀なのか？

大谷



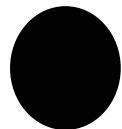
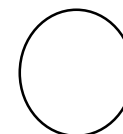
ベーブ・ルース



60点

得点をそのまま比べる

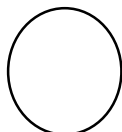
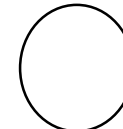
70点



50点

点数と平均点を比べる

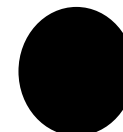
50点



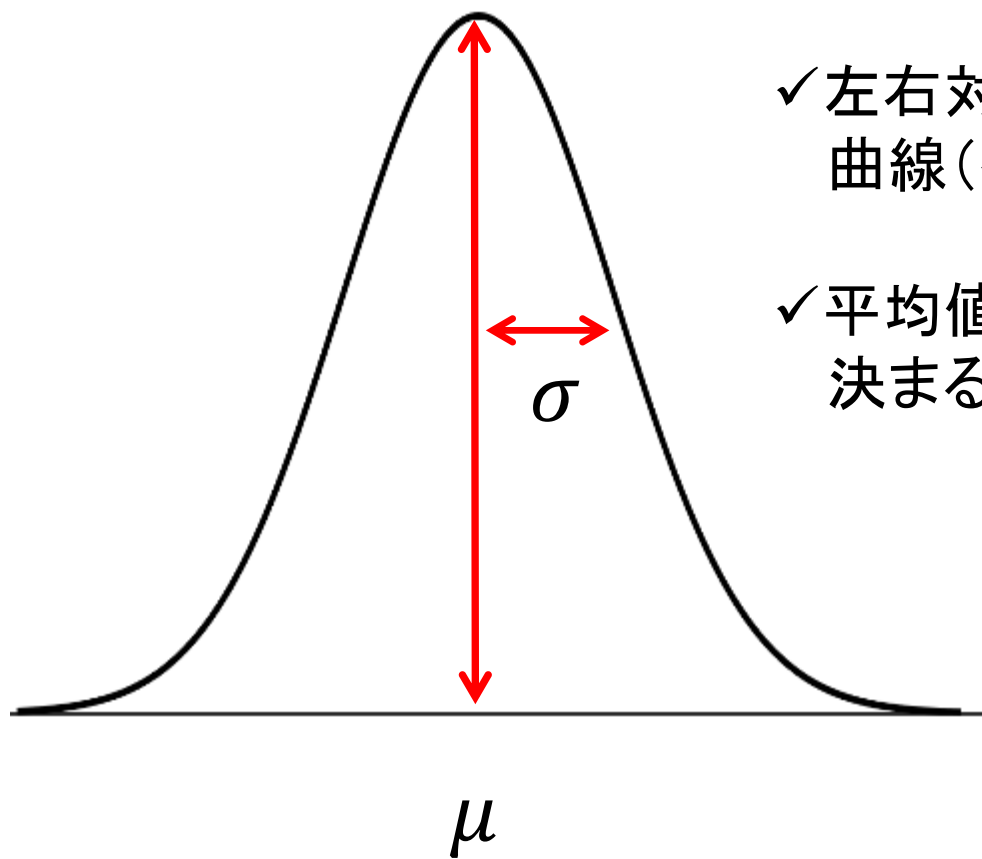
70

偏差値を比べる

63.3

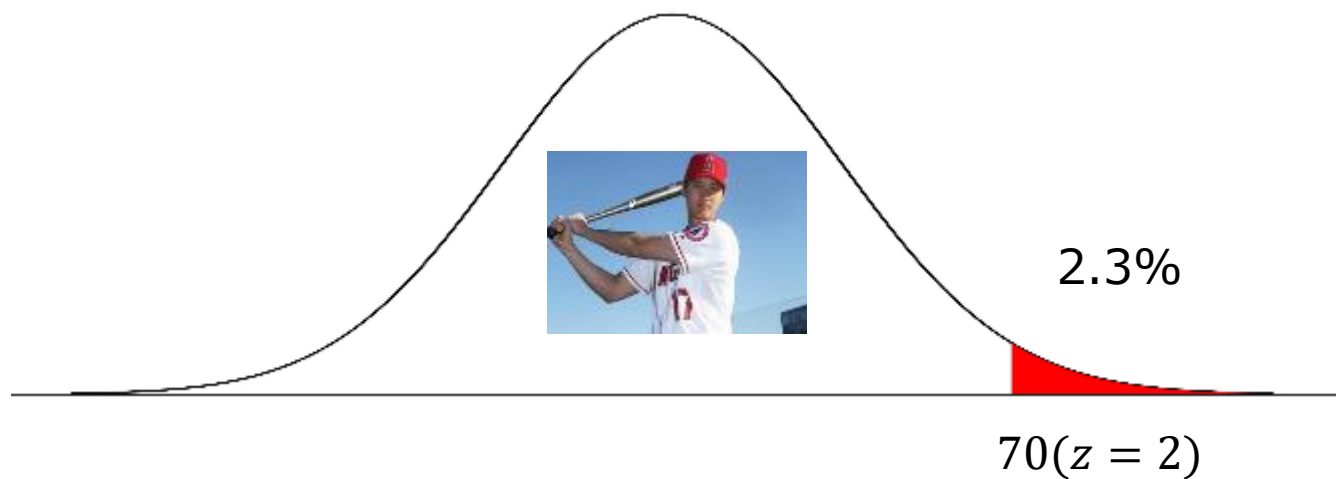
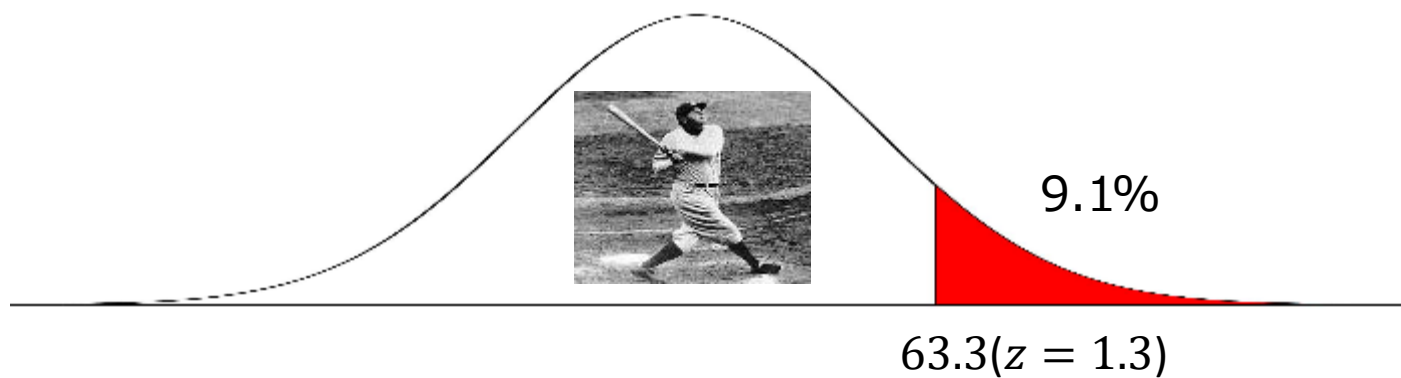


正規分布



- ✓ 左右対称になった西洋の釣鐘と似た形状の曲線（ベルカーブ）
- ✓ 平均値 μ 、標準偏差 σ の2つのパラメータが決まると形が決まる。

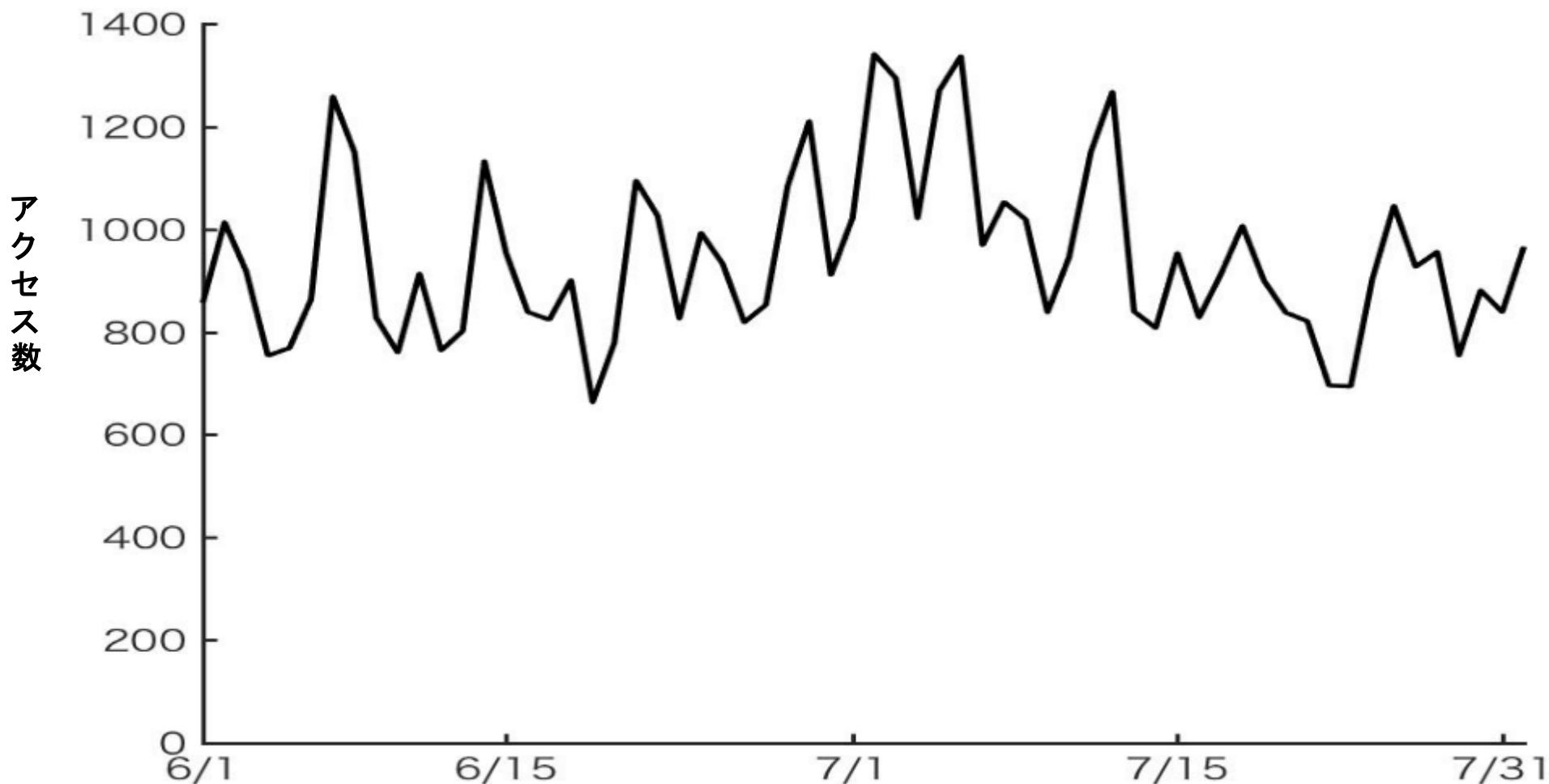
正規分布



移動平均法

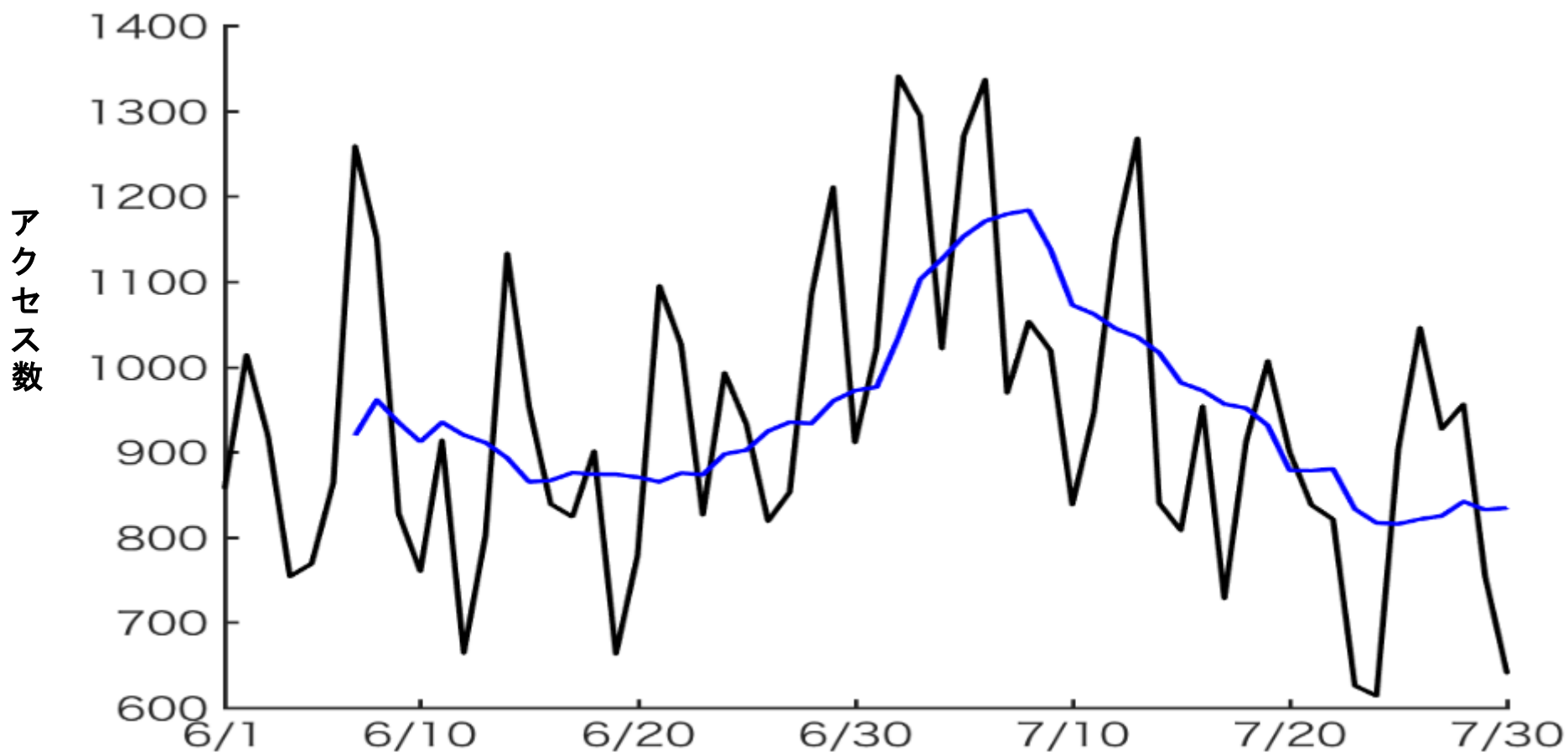
移動平均法を使ったトレンドの抽出

課題：「アクセス数のトレンドを推定せよ」



移動平均法を使ったトレンドの抽出

・ 時系列データ = **トレンド** + 周期変動 + 不規則変動



ベースラインの推定

