プロセス思考教室

データ分析研修

株式会社カレントカラー

Current Color

目的とゴール

- 目的
  - 業務改革を成功に導く

- 今回のゴール
  - データ分析の重要性を知る
  - データ分析の進め方を知る
  - グラフの選択と読み取り方のポイントを知る
  - バラツキに着目する意義を心に刻む

目次

- 1. データ分析の進め方
- 2. グラフ化のポイント
- 3. バラツキに着目する
- 4. バラツキを定量化する

参考資料

- 正規性の検定
- 数の起源、因果関係と相関関係、偏差値等

アジェンダ

名称	データ分析研修
時間・場所	
定員	８名
目的	業務プロセス改革を成功させる
今回のゴール	データ分析の重要性と、データ採取・グラフ化・分析の進め方を知る データを味方につけて、改革プロジェクトを**確実に進められる**ようになる

議題	担当	進行目安		時間
オープニング・悩み事の共有	全員			10分
１．データ分析の進め方	講師			15分
２．グラフ化のポイント	講師			35分
３．バラツキに着目する	講師			30分
４．バラツキを定量化する	講師			20分
クロージング・気づきの共有	全員			10分

※休憩は ありません。

１．データ分析の進め方

データ = 事実

データをもとに考え 認知バイアスを避ける

データは雄弁

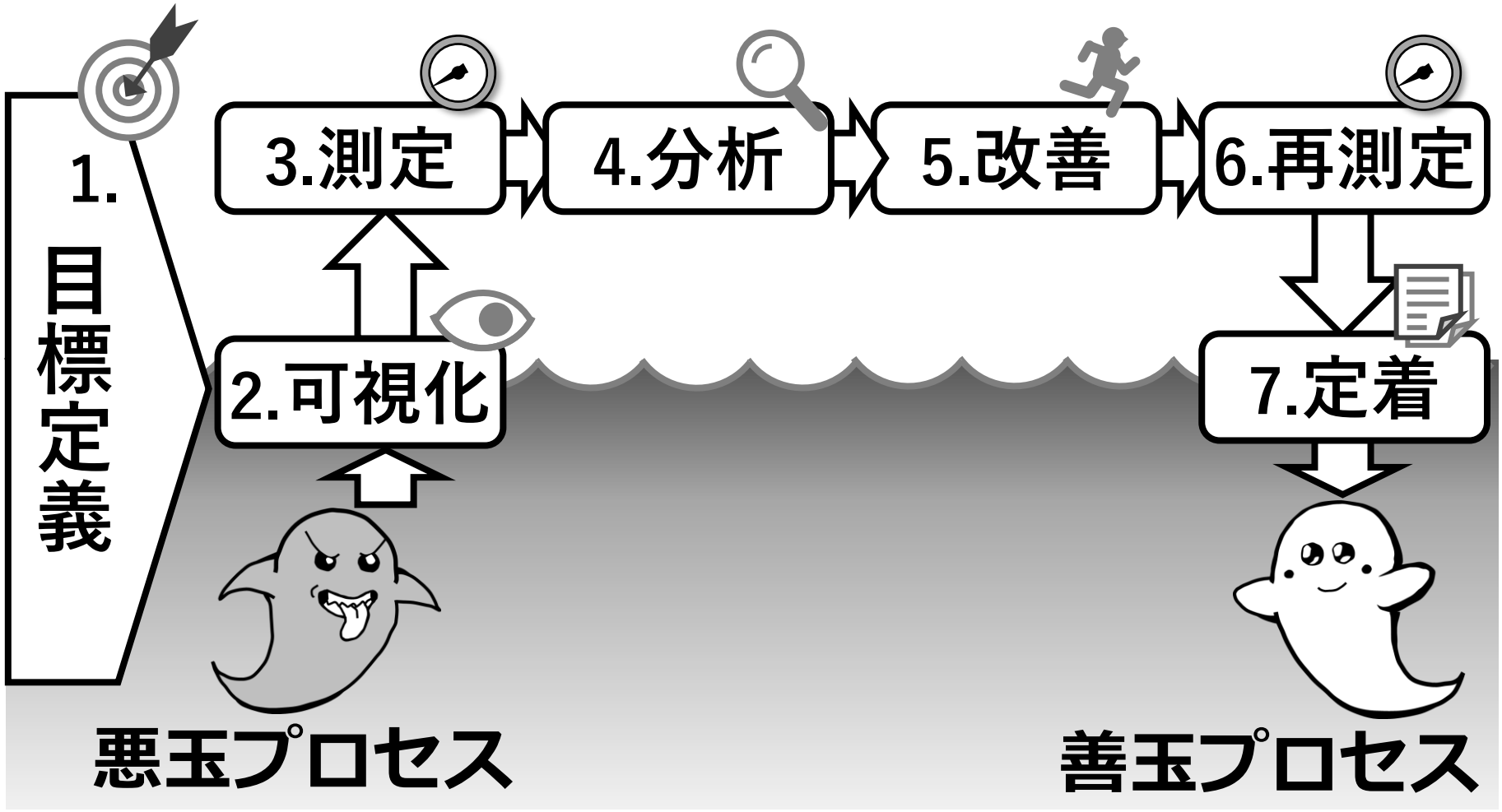
データは…

- 現実を浮き彫りにする
- 多角的に分析できる
- 共通認識を形作る
- 人を動かす！

改善７ステップ

目 視 測 分 改 測 着

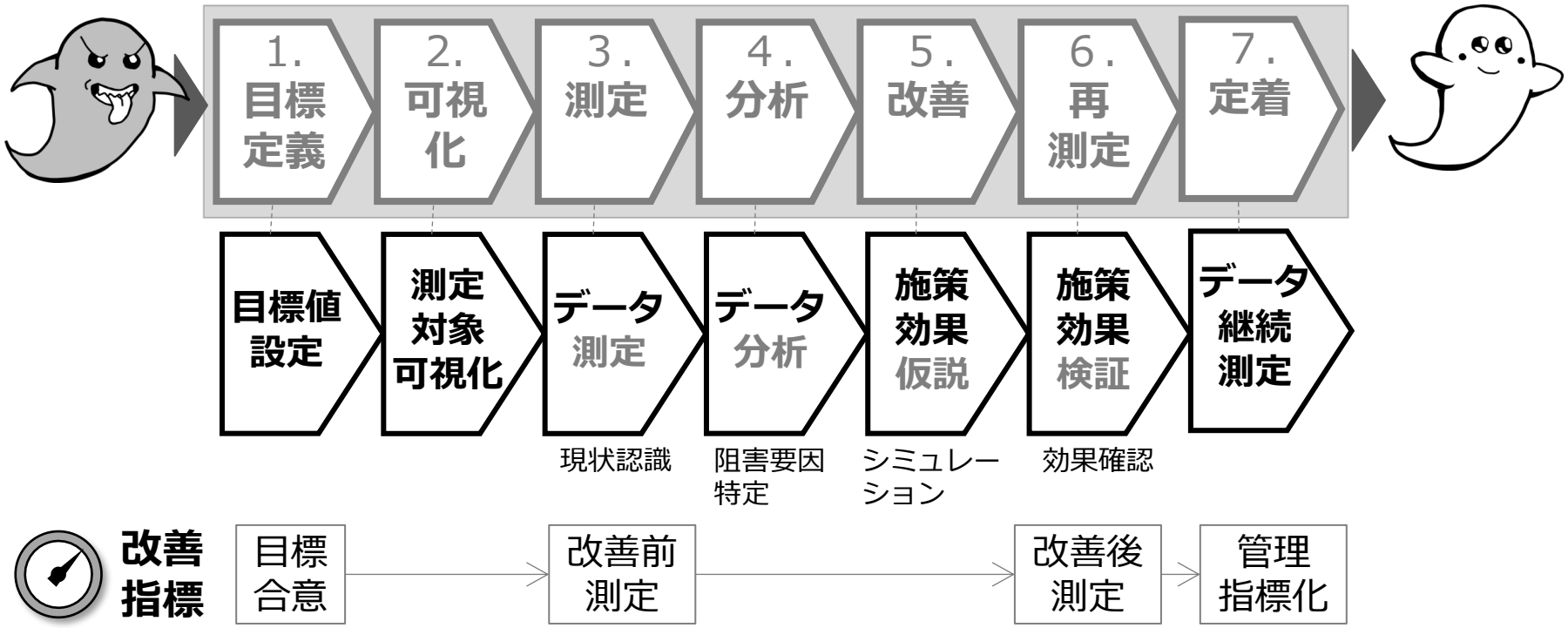
あらゆる改善・改革活動に使える改革手法



データ分析プロセス

目 視 測 分 改 測 着

データ分析活動＝改善7ステップ

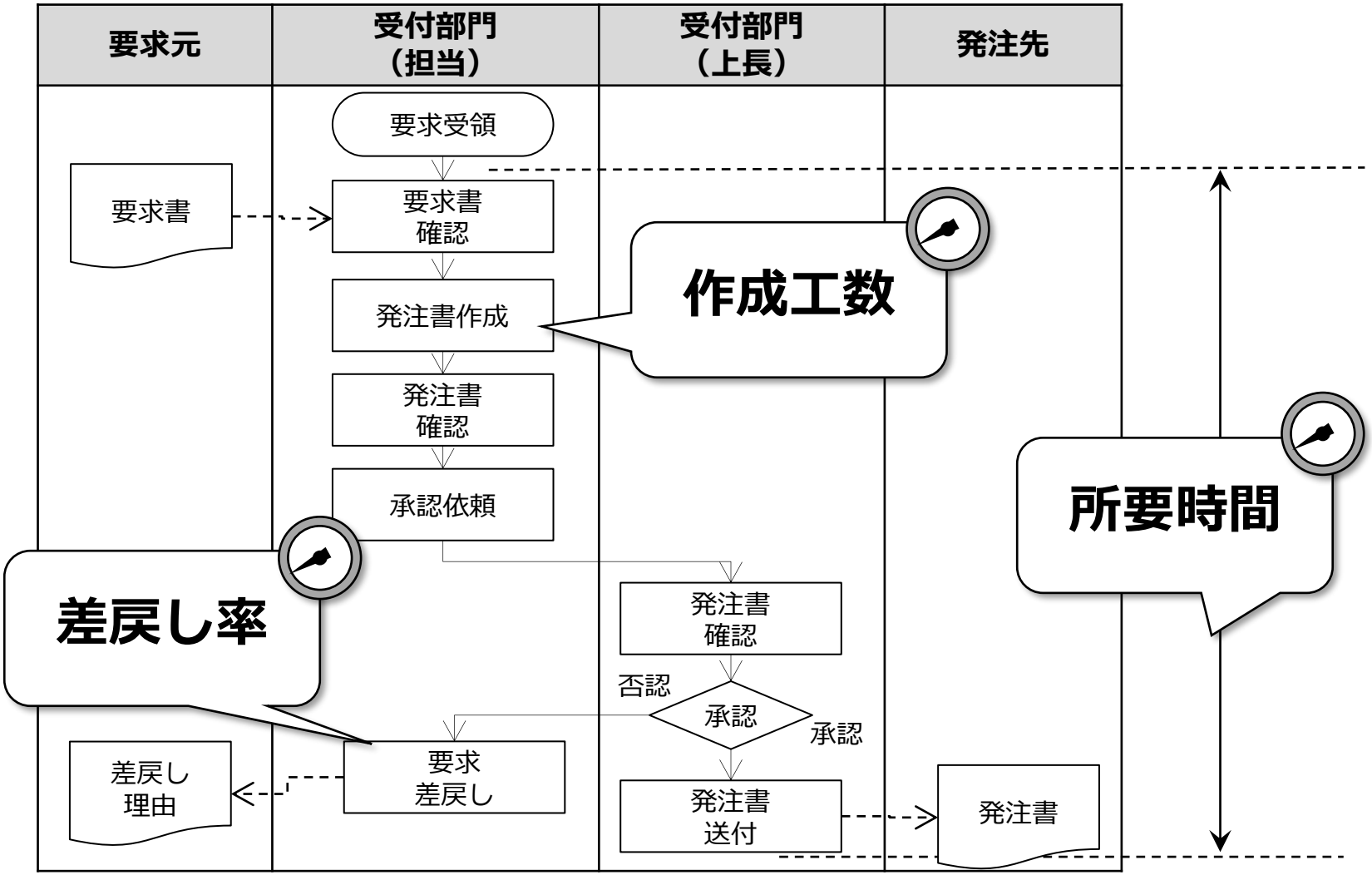


データ分析を実施する時点で、
「1. 目標定義」…**何のために**測定しているのか
「2. 可視化」…**何を**測定しているのか（対象プロセスは何か）
が明確になっていること。

プロセスの可視化

目 視 測 分 改 測 着

プロセスに紐づけ データの意味を正確に定義



データ収集

- データ収集にもコストがかかる
  - データの登録・集計・クレンジング・グラフ化もタダではない
  - 非公開のベンチマークデータは有償
- 仮説の検証に必要十分なデータを集める
  - 現象に対して原因が見つかる精度
  - スコープ内のデータはMECEに収集・分析
- 最初から詳細なデータを集め過ぎない
  - アタリをつけてから、深堀りする

ほぼ**自動**で大量のデータが集められるなら、人工知能に相関関係を発掘させてもいい。そこから新たな知見が得られることもある。一方、**手作業**で限られた範囲からデータを集めるなら、データ採取の目的を明確にし、測定や分析の手間を最小限にし、立てた仮説を検証し、改善活動に繋げられるように計画しよう。

データ収集

目 視 測 分 改 測 着

集計や分析の自由度のため、データ収集は**レコード形式**で

×

3月度工数							
A部門					B部門		
氏名	大区分	小区分	工数		氏名	大区分	小区分	工数
本田曜子	直接工数	伝票処理	30		島田真理子	直接工数	情報分析	60
		情報分析	40				問合対応	30
		問合対応	20			間接工数	会議	15
	間接工数	会議	10					
		事務処理	5		猪俣浩二	直接工数	情報分析	80
							問合対応	5
西山権左衛門	直接工数	伝票処理	80			間接工数	会議	15
	間接工数	会議	20				事務処理	3
		事務処理	1					

○

集計月	部門	氏名	大区分	小区分	工数
3月	A部門	本田曜子	直接工数	伝票処理	30
3月	A部門	本田曜子	直接工数	情報分析	40
3月	A部門	本田曜子	直接工数	問合対応	20
3月	A部門	本田曜子	間接工数	会議	10
3月	A部門	本田曜子	間接工数	事務処理	5
3月	A部門	西山権左衛門	直接工数	伝票処理	80
3月	A部門	西山権左衛門	間接工数	会議	20
3月	A部門	西山権左衛門	間接工数	事務処理	1
3月	B部門	島田真理子	直接工数	情報分析	60
3月	B部門	島田真理子	直接工数	問合対応	30
3月	B部門	島田真理子	間接工数	会議	15
3月	B部門	猪俣浩二	直接工数	情報分析	80
3月	B部門	猪俣浩二	直接工数	問合対応	5
3月	B部門	猪俣浩二	間接工数	会議	15
3月	B部門	猪俣浩二	間接工数	事務処理	3

→ 1レコード

データ分析

目 視 測 **分** 改 測 着

観点	主なグラフ	主な分析手法
比較 （差異の把握）	• 棒グラフ • 積み上げ棒グラフ • レーダーチャート	• ソート • 層別（ヒートマップ）
構成比 （部分の割合）	• 円グラフ • 積み上げ棒グラフ • パレート図	• ABC分析 • パレート分析
分布 （広がり・偏り）	• ヒストグラム • 箱ひげ図 • パレート図	• 分散・標準偏差分析 • 正規性検定
時系列 （変化・傾向）	• 折れ線グラフ • 管理図	• 特別要因分析 • 移動平均
相関 （変数間の関係）	• 散布図 • バブルチャート	• 相関分析・回帰分析 • データマイニング （アソシエーション分析） （バスケット分析） • クロス集計分析 • ディシジョンツリー
分類 （概念の構造）	• レーダーチャート • バブルチャート	• クラスター分析 • KJ法（親和図法）

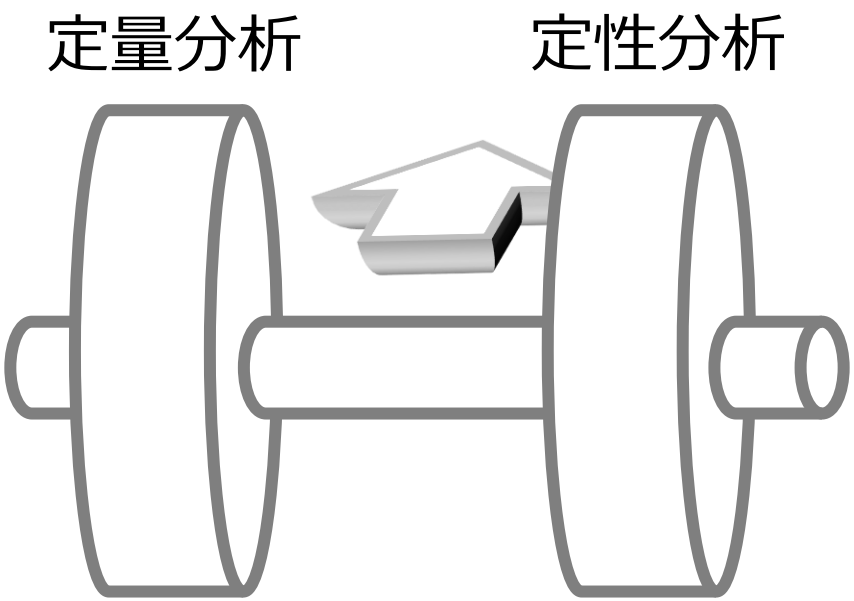
データ分析

目 視 測 **分** 改 測 着

観点	分析手法	説明
構成比	• **ABC分析**	全体に占める割合でランク分けし対策を決める Aランク：上位70% を生み出す 数少ない要素 Bランク：上位70%～95% までを 生み出す要素 Cランク：残り5% を生み出す多数の要素 （ロングテール）
分類	• **クラスター分析** • **KJ法（親和図法）**	性質や量の近いものをグルーピングする
相関	• **相関分析** • **データマイニング** （アソシエーション分析） （バスケット分析）	何と何が強い相関を持つかを探る （例）オムツとビールはセットで売れる （例）健康の本を買った人にサプリもオススメ
相関	• **クロス集計分析**	どの値と属性に強い相関があるかを探る （例）アンケートの結果と、回答者の属性 （例）居住地と、政策への評価
時系列	• **移動平均**	時系列データから、一定の範囲の平均値を取り、その範囲を移動させて、全体を平準化する手法。 不規則な変動を除去し、傾向を見やすくする。

定量 vs. 定性

定量分析と定性分析は、車軸の両輪



- 定量分析は定性情報で、定性分析は定量情報で、裏付けを取る
- アンケートの自由記述などの定性情報は、
  親和図（**KJ法**）などでグルーピングして定量的に把握してみる。
- 影響度は「声の大きさではなく、**声の多さ**で評価する」

定量 vs. 定性

定量情報と定性情報は、使い分け、組み合わせる

観点		定量情報	定性情報
表現方法		数値＋単位、グラフ	自然言語、イメージ図
長所・短所	表現の自由度	× 低い	○ 高い
	複雑な内容の伝達	× 困難	○ 容易
	誤伝達や誤解のリスク	○ 小さい	× 大きい
	集計や比較の手間	○ 小さい	× 大きい
	理解・解釈までの時間	○ 速い	× 遅い
好適領域（例）		客観的事実、統計的記述 **実績**のアピール、財務	主観的内容、心理的状態 **意味・意義**の主張、価値
収集方法（例）		選択肢 装置による計測	自由記述 インタビュー

- 定量情報も定性情報も「コミュニケーションツール」
  目的に応じて適切に使い分ける
- 様々な価値観を持つ不特定多数の人に訴求するには定量情報が有効

2. グラフ化のポイント

データの魅せる化

魅せる化

生データの羅列から
傾向や判断基準を
読み取れるのは、天才だけ

▼

見せ方の工夫＝**魅せる化**
が必要

魅せる化

Insight!

「魅せる化」とは？

↓

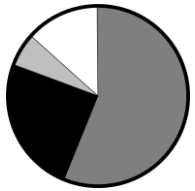
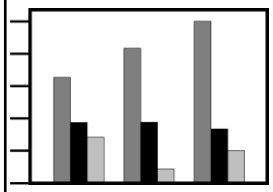
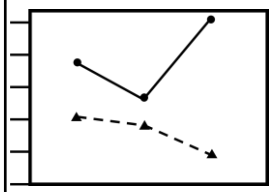
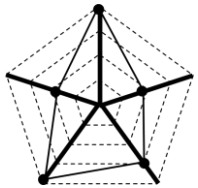
Action!

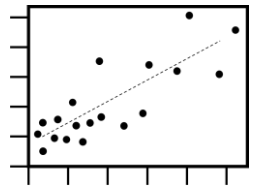
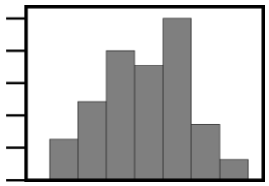
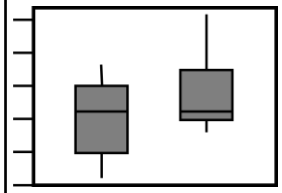
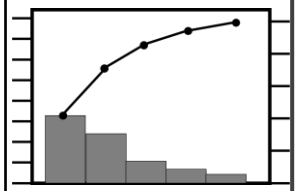
データの表現方法の工夫により
人を**改善行動に駆り立てる**こと

- 適切な表現・グラフの選択
- 適切なスケール・粒度
- 適切な貼り出し場所
- 適切な更新頻度

グラフ化

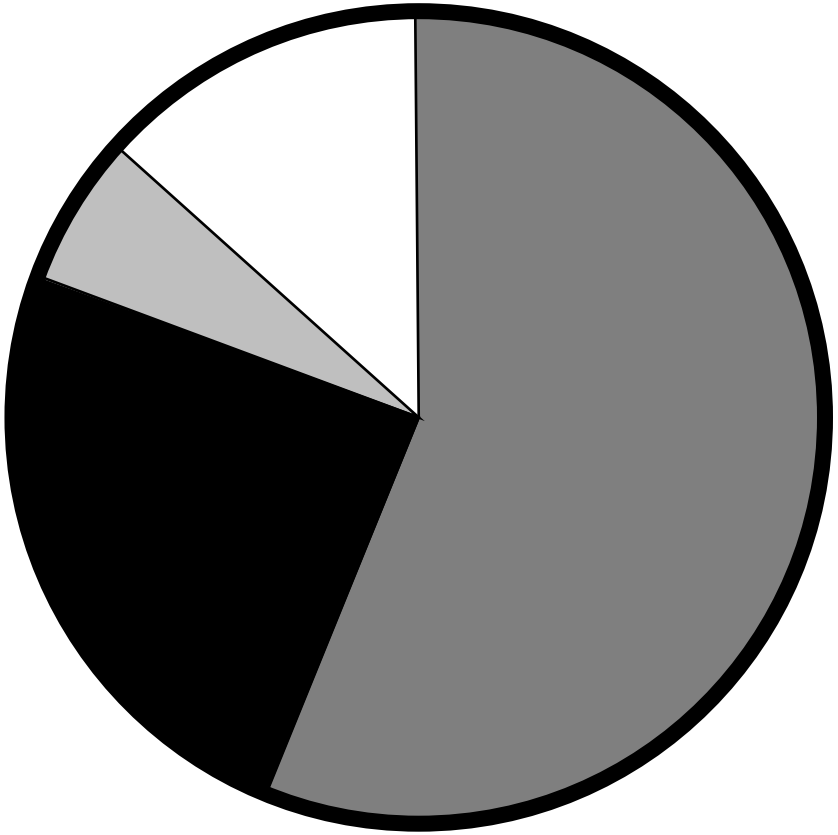
定量データを収集したら 先ず グラフ化 しよう！

観点	グラフ	形状
構成比	円グラフ	
比較	棒グラフ	
時系列	折れ線グラフ	
分類	レーダーチャート	

観点	グラフ	形状
相関	散布図	
分布	ヒストグラム	
分布	箱ひげ図	
分布	パレート図	

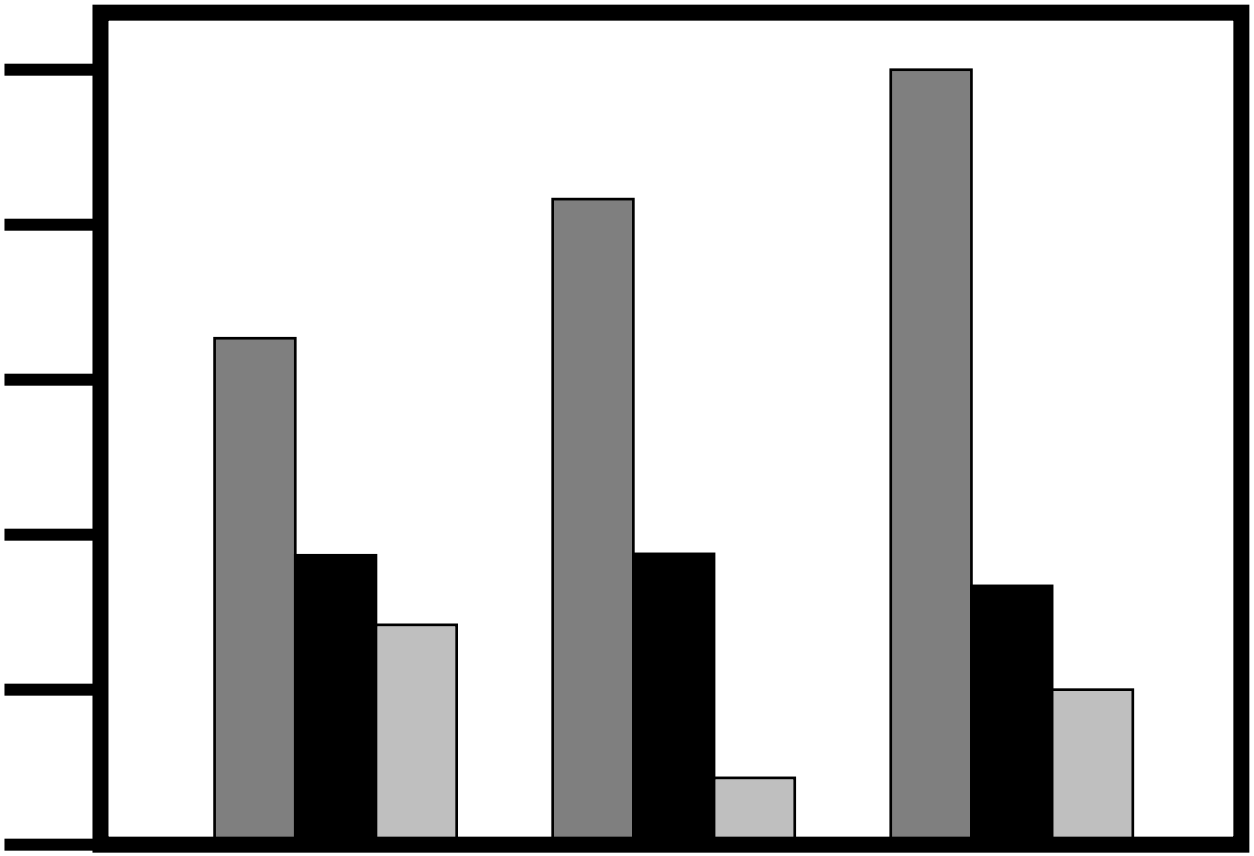
円グラフ pie chart

- **全体の中の割合**を示すには、円グラフを使う。
- 精密な比較ではなく、**大まかな比率**を掴むのに向いている。



棒グラフ bar chart

- ２つ以上の値を**比較**する際に使われる。
- 円グラフよりも**微妙な差異**を見分けられる。

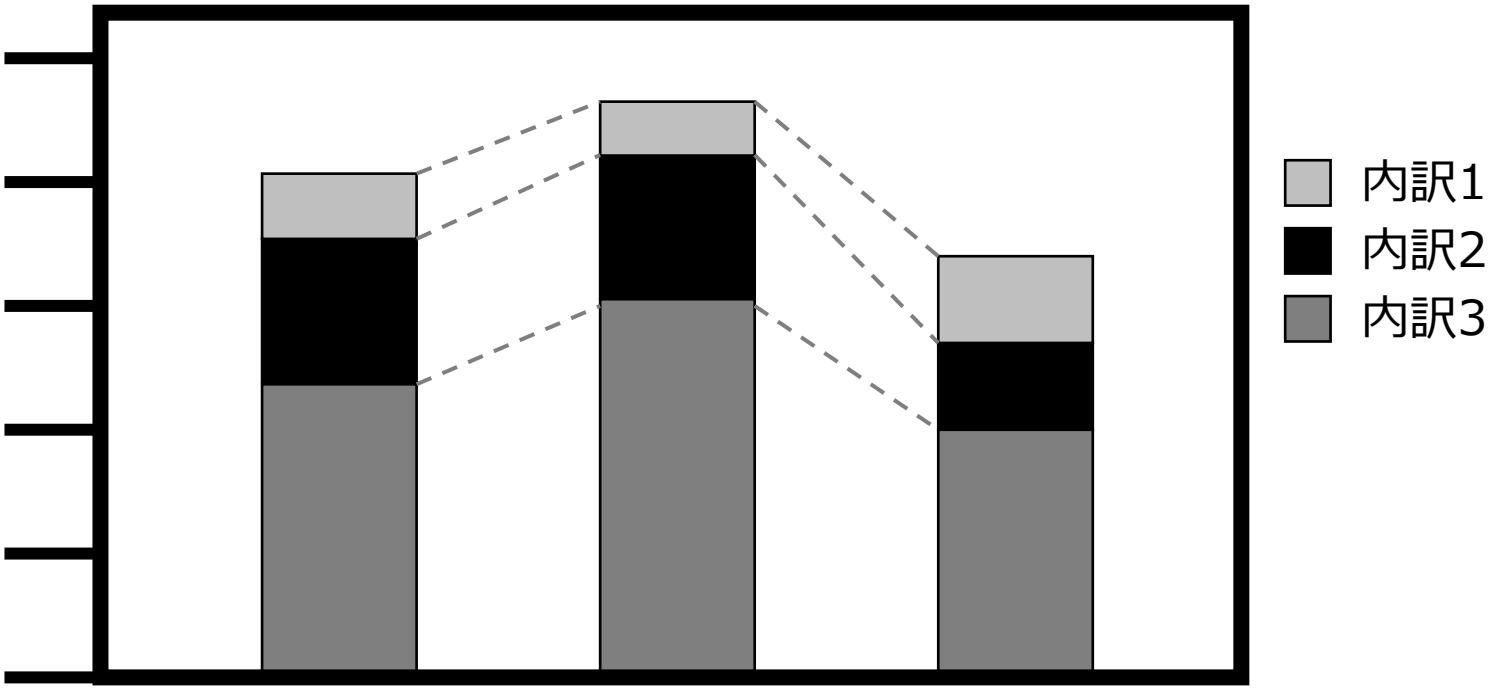


積み上げ棒グラフ

構成比 | 比較

stacked bar chart

- **要素**同士の比較も、**内訳**同士の比較もできる
- **要素**に対する**内訳**の割合も分かる

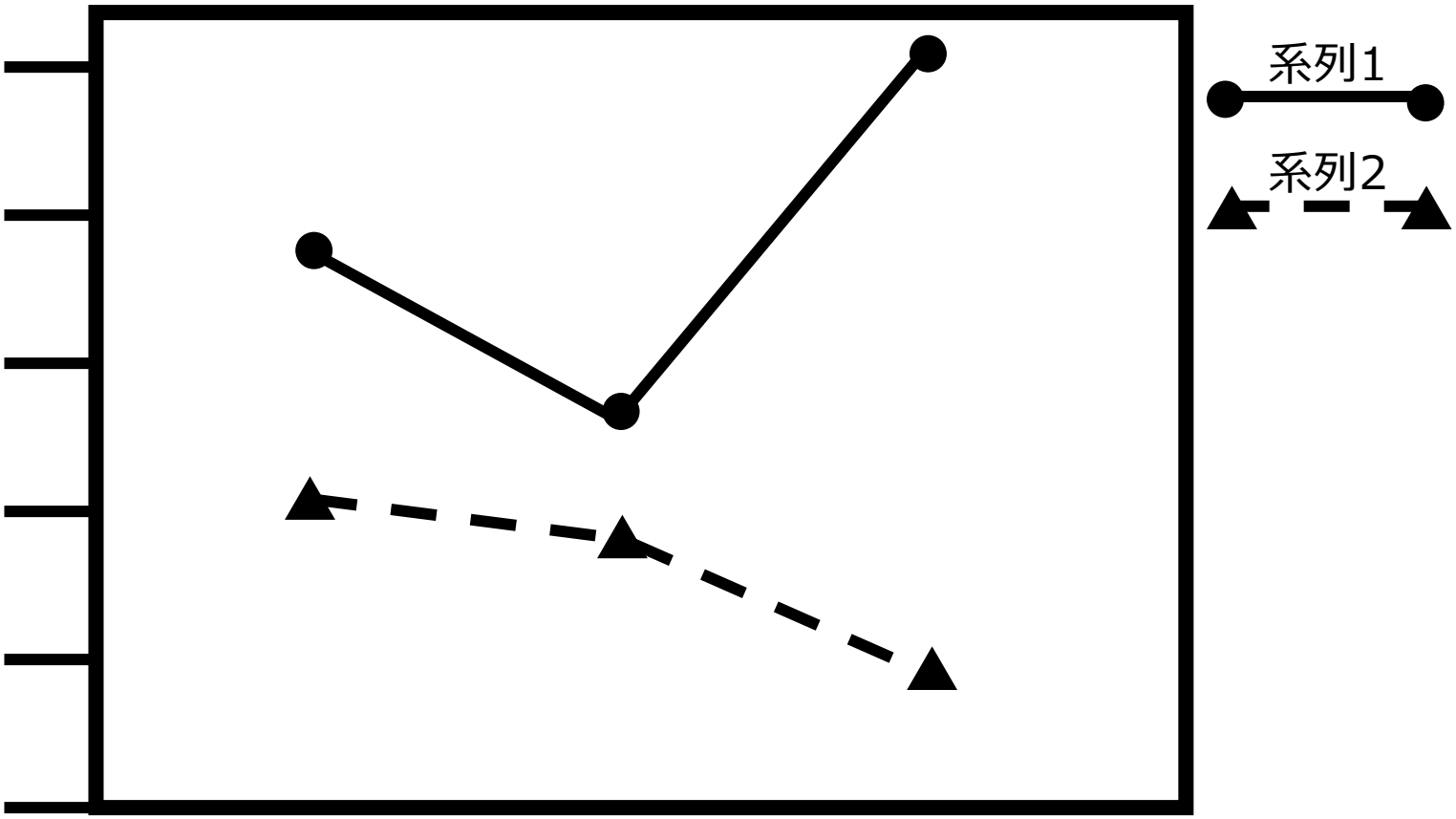


（例）作業区分ごとの工数
（例）売上高や原価の内訳
（例）アンケートの満足度の内訳

→ 比率の比較・変化

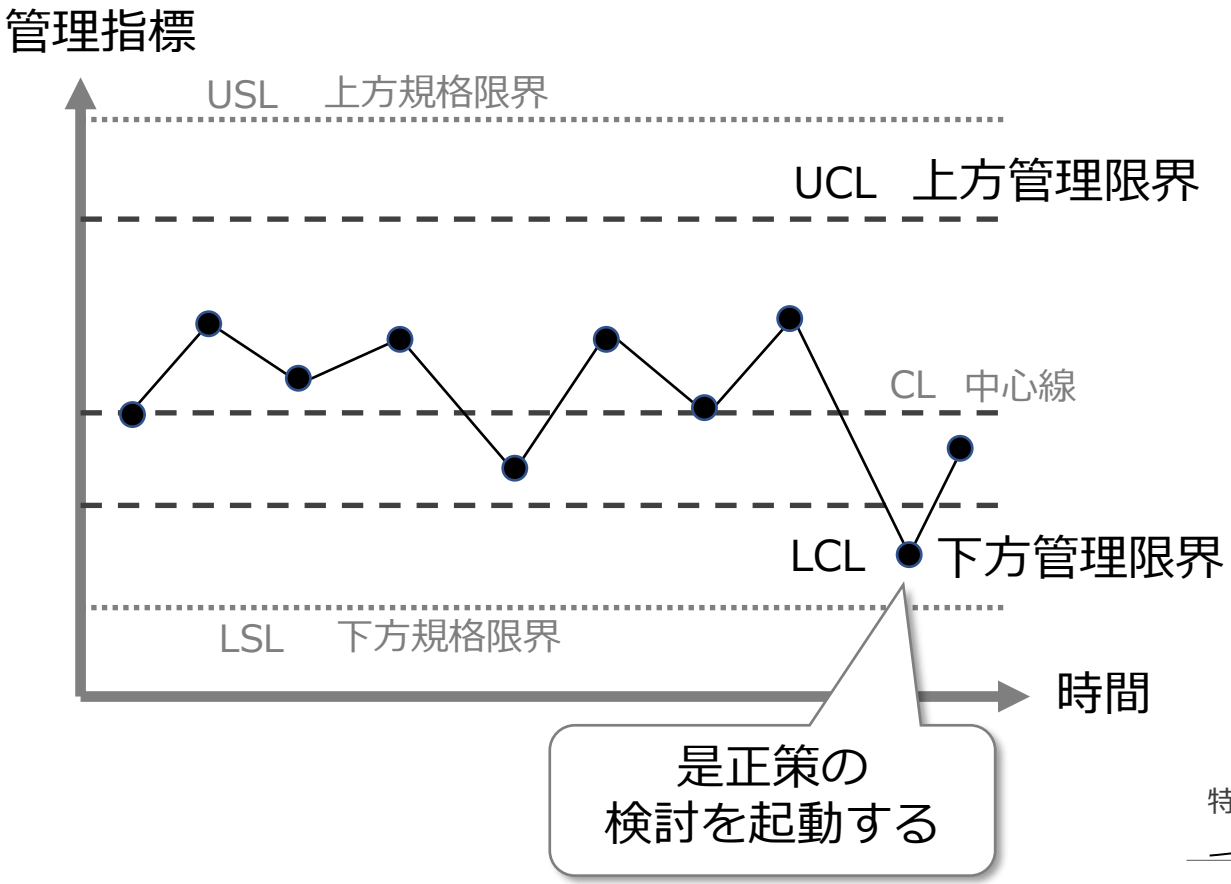
折れ線グラフ line chart

- 時間の経過で数量がどのように**変化**するかを表わす
- 線分の**傾き**で、変化の急激さ・緩やかさが分かる



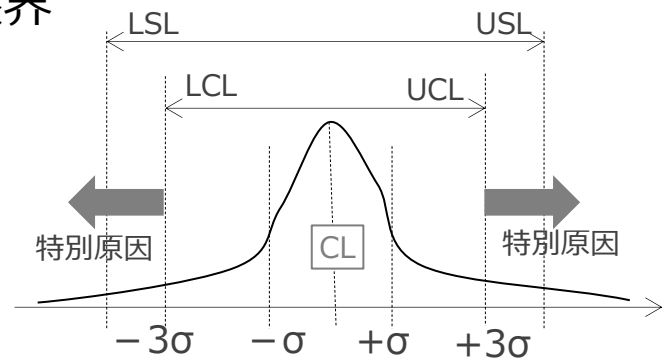
管理図 control chart

- **時系列**データで、品質や工程の**安定性**を判断する
- 調査すべき異常、**特別原因**を見つける



お客様にとっての限界
USL = Upper Spec Limit
LSL = Lower Spec Limit
内部管理の限界
UCL = Upper Control Limit
（平均＋３標準偏差）
LCL = Lower Control Limit
（平均－３標準偏差）

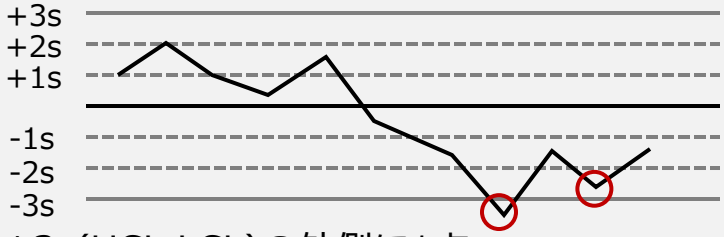
・定義上、「UCL-LCL」は「USL-LSL」の内側にあるべき。
・慣例として、±３標準偏差を超える稀な事象を「特別原因」と呼ぶ。



管理図

管理図から「特別要因」を
どうやって見つけ出すの？

例えばこんなルールがあるよ。



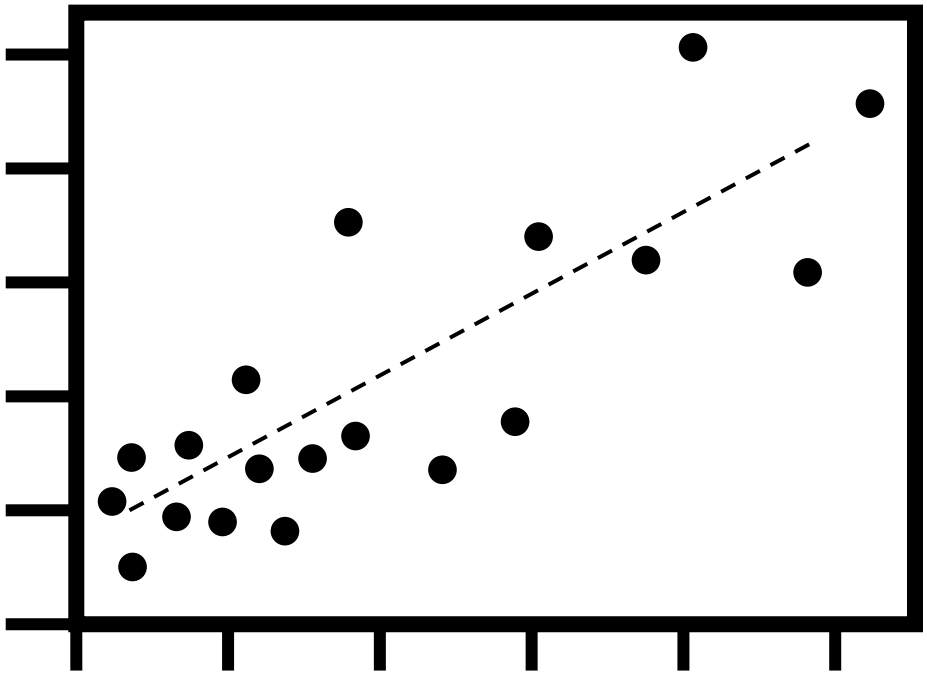
1. ±3s(UCL,LCL)の外側に1点
2. ±1sの外側（同じ側）に連続９点
3. 連続6点で上昇または下降
4. 連続14点で上昇下降を交互に繰り返す
5. 連続3点のうち±2sの外側に2点
6. 連続5点のうち±1sの外側に4点
7. 連続15点が±1sの内側（同じ側）
8. 連続8点が±1sの外側（同じ側）

統計的な４つの判断に、シューハート博士
が考案した４つの判断が加わっていて…

……

散布図 scatter plot

- ２つの連続量の間に、**相関関係**があるかを視覚化する。
- データ群が右上がり傾向なら、正の相関がある、と言う。

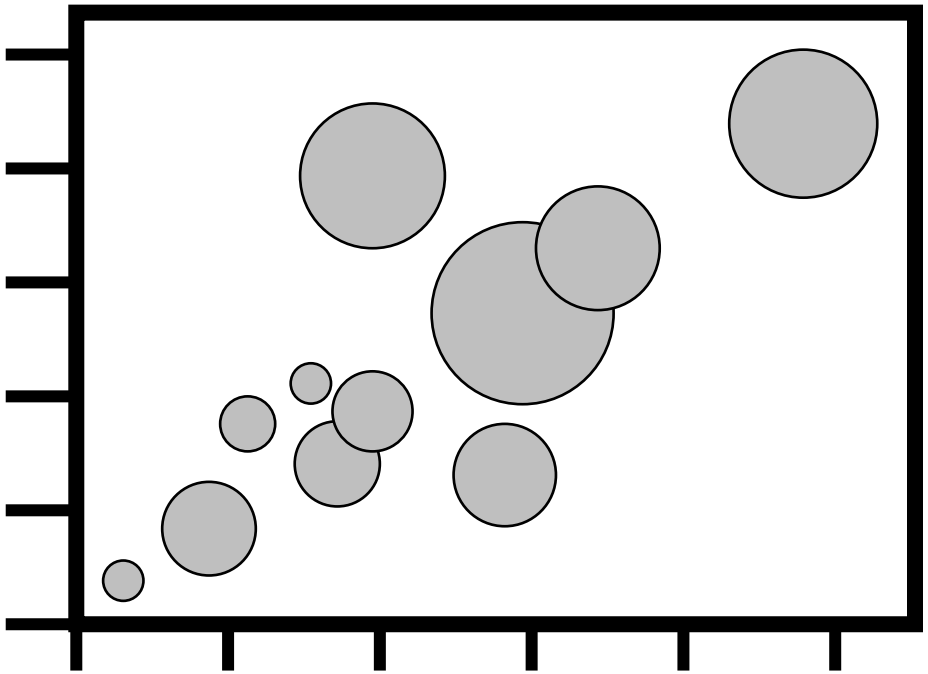


注意：**相関関係**があっても、それが**因果関係**であるとは限らない。
（例）残業が増えると、問合せ件数も増える。
従って、問合せ件数が増える原因は、残業にある。　→（×誤り）
（例）アイスクリームが売れると、電力消費量が増える。
従って、電力消費量が増えるのは、アイスクリームのせいだ。　→（×誤り）

分類 相関

バブルチャート bubble chart

- ３つの値の関係を、一つのグラフで視覚化する
- ２量の相関関係に加えて、もう１つの量も分析できる



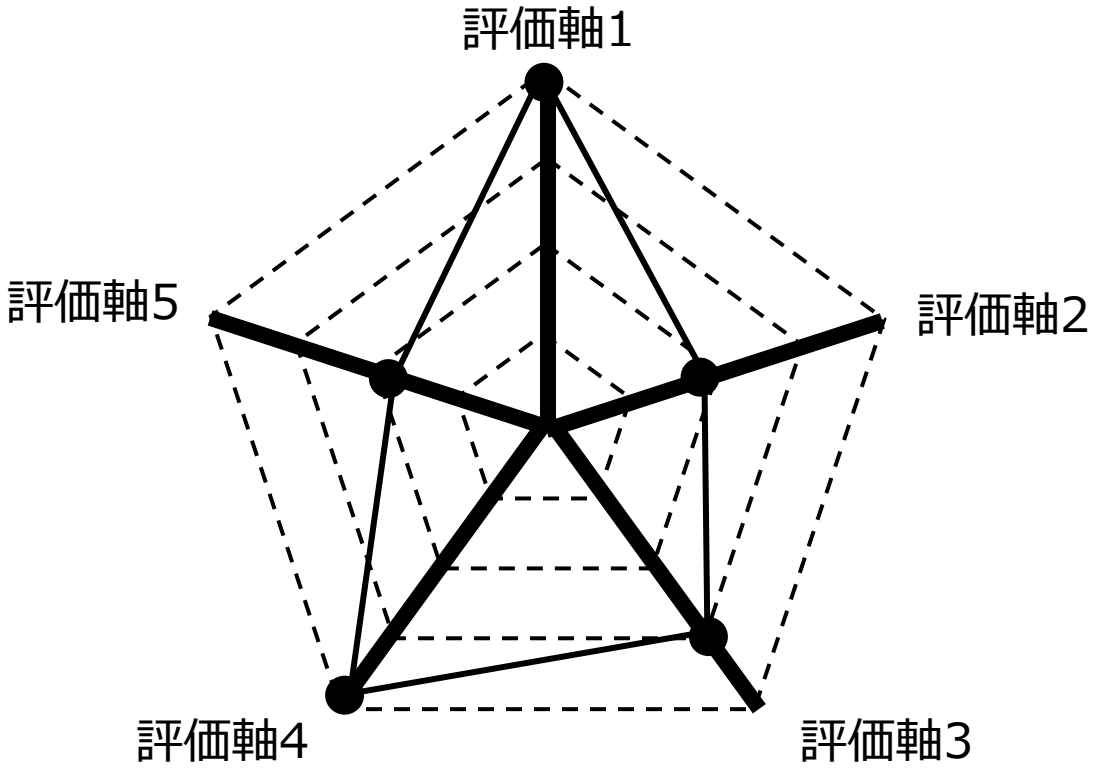
（例）各業界について、**縦軸**：売上高、**横軸**：利益率、**半径**：会社数
（例）各商品について、**縦軸**：成長率、**横軸**：占有率、**半径**：売上高（PPM分析）

注意：バブルが多くなりすぎると、重なって見づらくなる。
　　バブルの大きさに余り差が無い場合は、比較が難しい。

比較　分類

レーダーチャート radar chart

- 複数の評価軸の**バランス**を視覚化する
- 意味の近い評価軸を近くに並べると、傾向が分かり易くなる

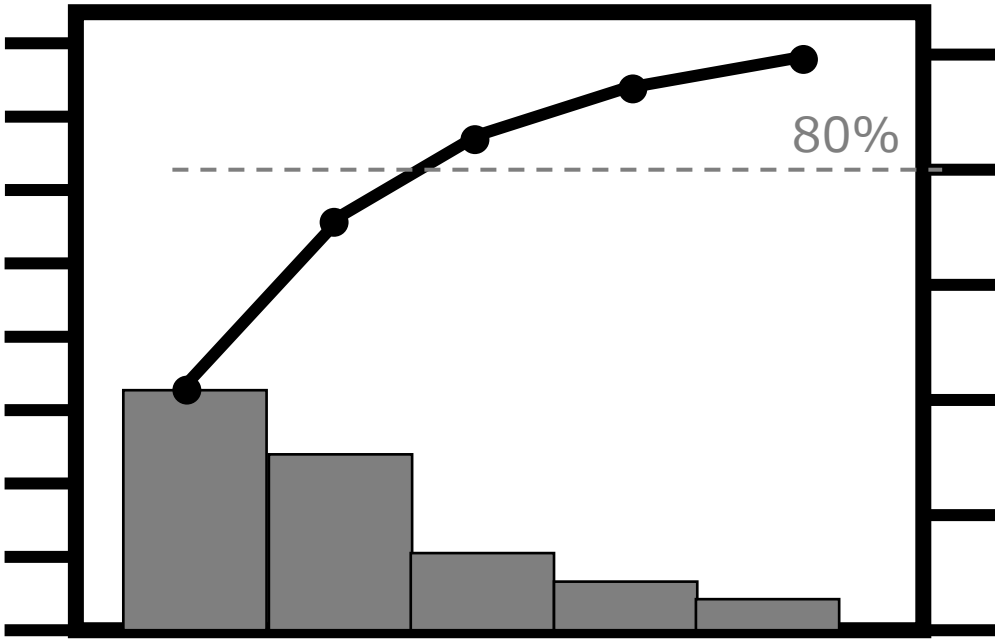


- 長所、短所が結線の形状で視覚的に分かる。

分布 | 構成比

パレート図 Pareto chart

- 上位の**少ない要因**が、全体に**大きく影響**を与えているということを、分かり易く示す。



パレートの法則：
全体の数値の大部分は、
全体を構成するうちの
一部の要因が生み出している

- いわゆる「２：８の法則」を視覚的に確認できる。
- 「なぜこの要因に絞って対策を立てているのか」と問われそうな箇所では、パレート図を用意する。

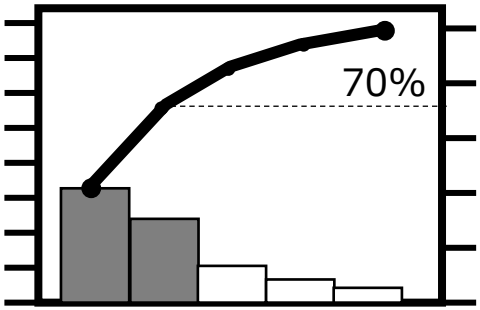
パレート図

分布　構成比

「パレート図」なんて
使ったこと無いなぁ…

たくさんの要因がある時に、
どこまで考慮すればいいのか
示せるよ。

そうか、
「この２つの要因で全体の70%を
　カバーしているから、**十分です**」とか
「この２つの要因だけでは全体の50%しか
　カバーしていないので、
　考慮範囲を**広げる必要**があります」とか
分かりやすく伝えられるわね！



3．バラツキに着目する

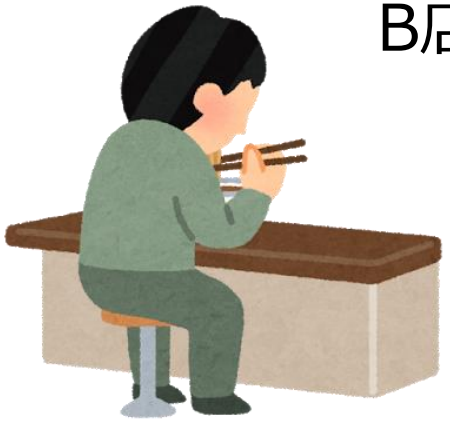
平均値を見るだけでは
データが泣いている

サービス品質とバラツキ

あなたなら、どちらのラーメン屋に入りますか？

お昼休みは、あと20分。

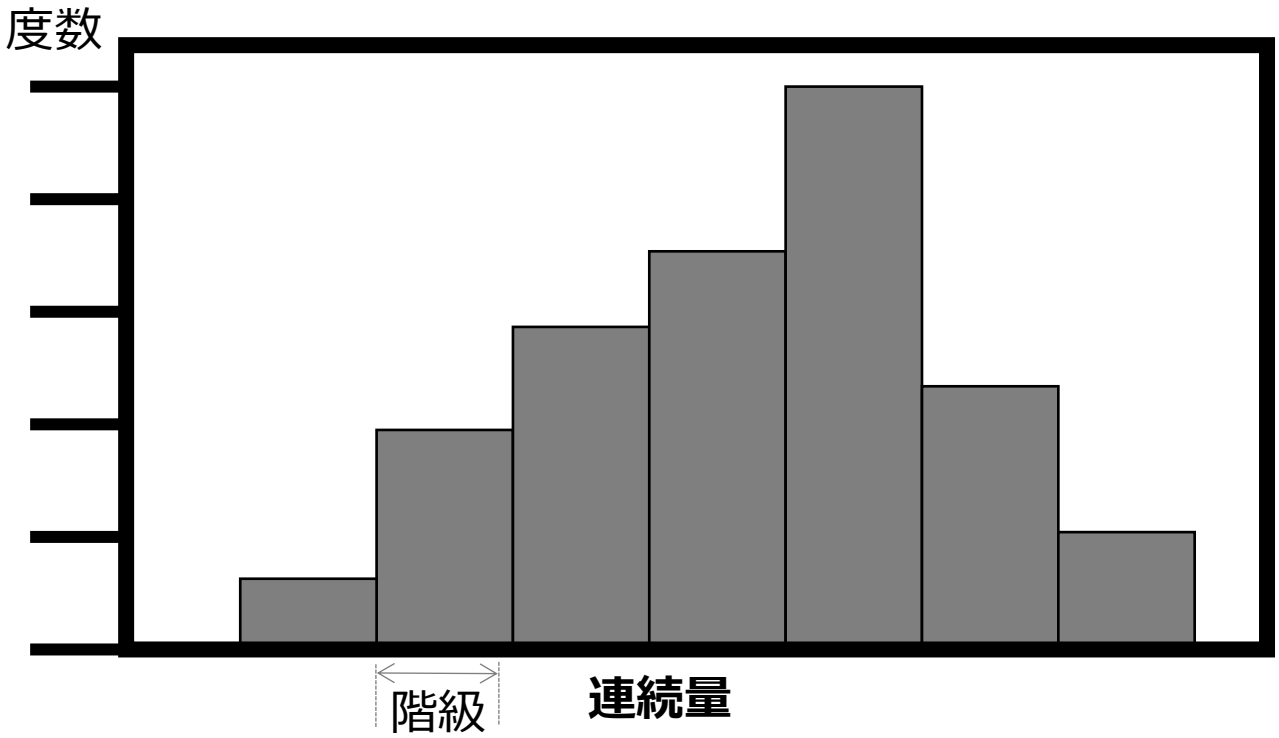
A店：注文してから、
平均すると７分で出てくる。
ほぼ確実に、６分から８分の
間に出てくる。



B店：注文してから、
平均すると５分で出てくる。
３分で出てくることもあれば、
20分かかることもある。

ヒストグラム histogram

- 「**分布の形状**」「**中心化傾向**」「**偏り**」が分かる
- **連続データ**は、まずヒストグラムで見てみる



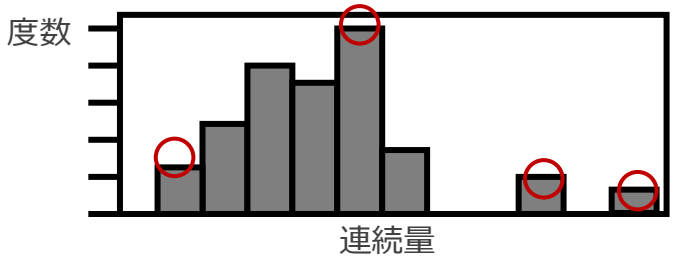
見た目は棒グラフと似ているけれど、それぞれの棒がくっついていて、
横方向に値が**連続**していることを表わしている。

ヒストグラム

連続値が得られたら「まずヒストグラム」
…っていうけど、そんなに大事なこと？

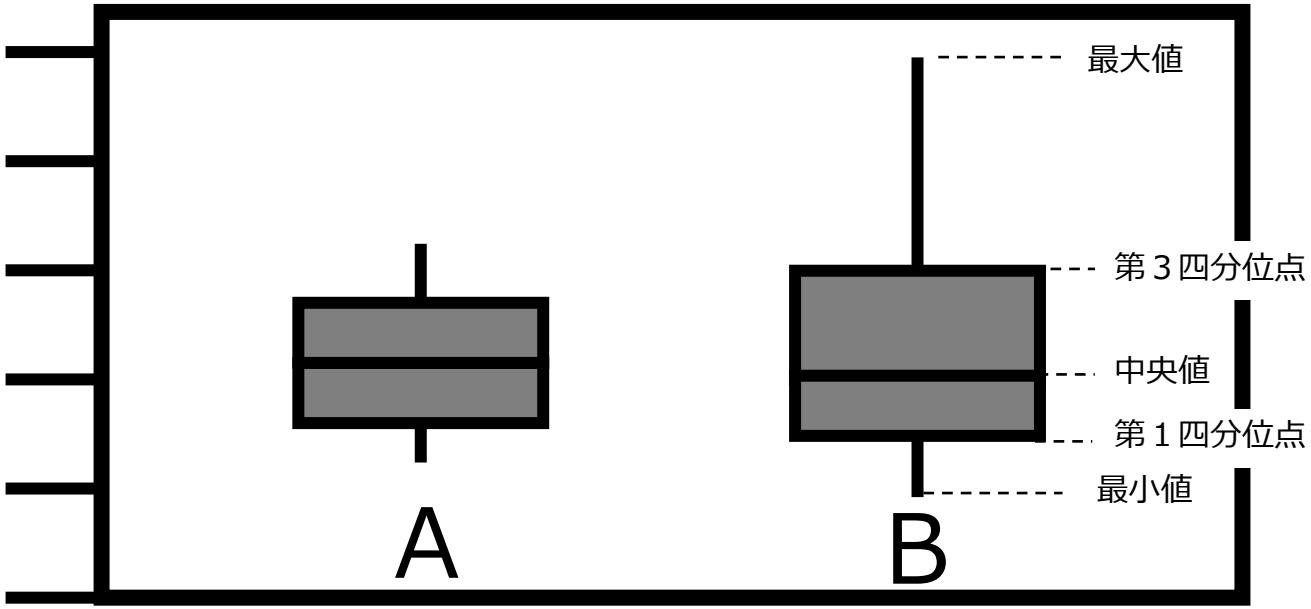
平均値・最大値・最小値だけでは、
「**バラツキ**具合」「**カタマリ**具合」が
分からないよ。もったいないよ！

確かに、
どのあたりが典型的な値なのか、
このあたりのバラツキの原因は何か、
とか、考えられるようになるわね！



箱ひげ図 box plot

- データの**バラツキ具合**を見るのに、良く使う。
- 複数の集団の**バラツキ具合を比較**するのにも便利。

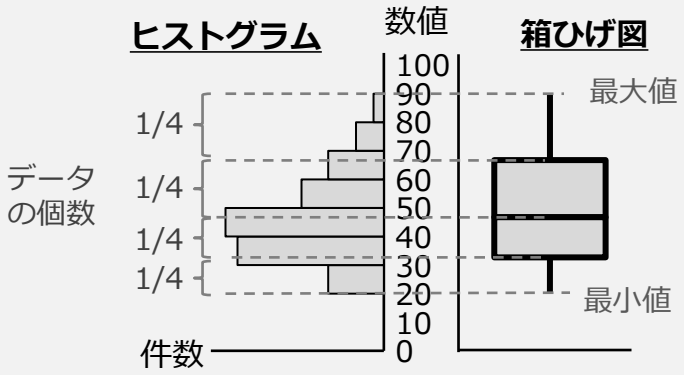


データを小さい順に並べた時、
最大値　　…最後の値
第３四分位点…７５%番目の値
中央値　　…５０%番目の値
第１四分位点…２５%番目の値
最小値　　…最初の値

箱ひげ図

「箱ひげ図」の見方が
良く分からない…

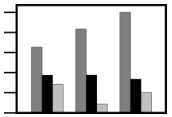
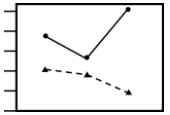
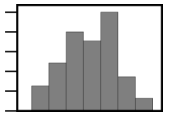
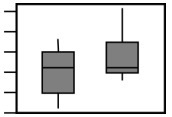
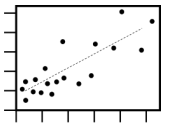
ヒストグラムを横にしたカンジだよ。



あ、見えてきた！
「箱」や「ひげ」が
長いほど「**バラついて**」いる、
短いほど「**カタマって**」いる、
…ってことね！

演習｜グラフの選択

以下の指標は、どんな**観点**に着目しているでしょうか。
また、どんな**グラフ**で「魅せる化」しますか。

A製品とB製品の顧客満足度のバラツキ	**円グラフ** 
直近６ヶ月の売上高の増減	**棒グラフ** 
日別の電話受付回数	**折れ線グラフ** 
残業時間の階級ごとの人数	**ヒストグラム** 
電子メール受信数と残業時間の関係	**箱ひげ図** 
欠陥数の原因別の比率	**相関図** 

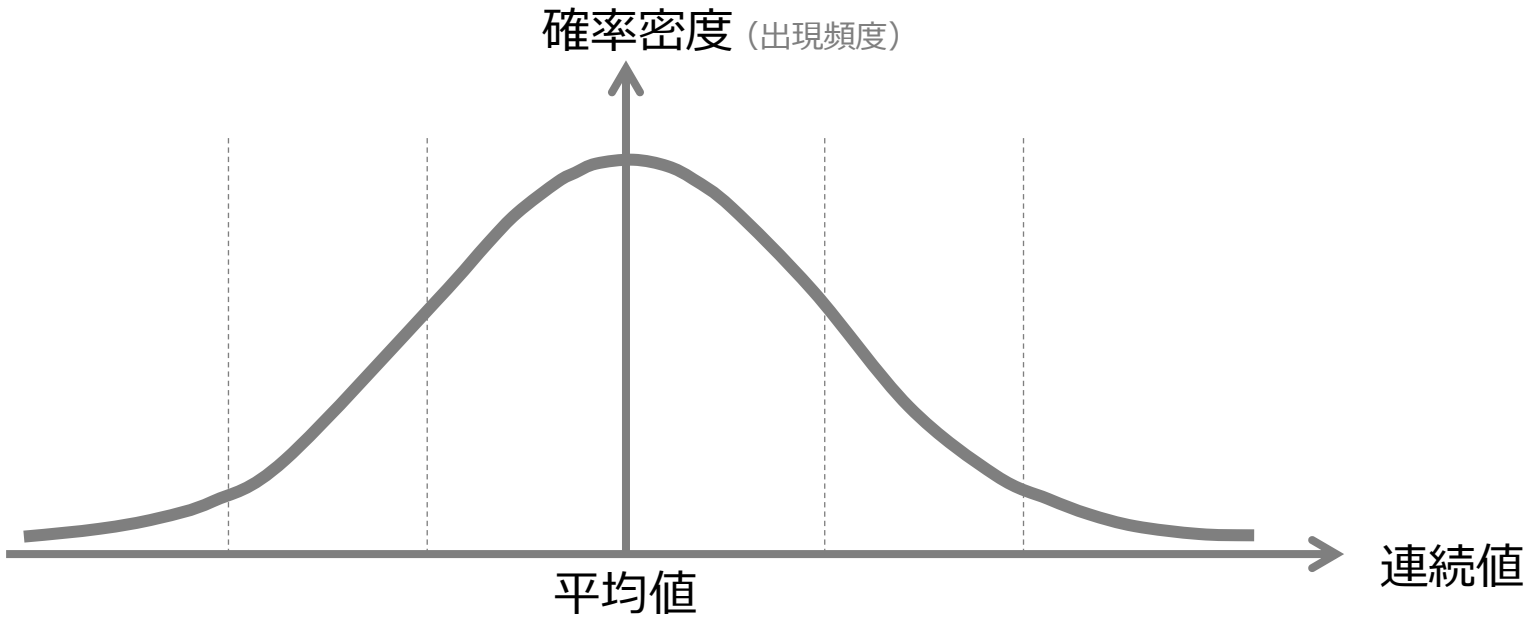
４．バラツキを定量化する

バラツキの大きさを
数値で表現する

正規分布とは

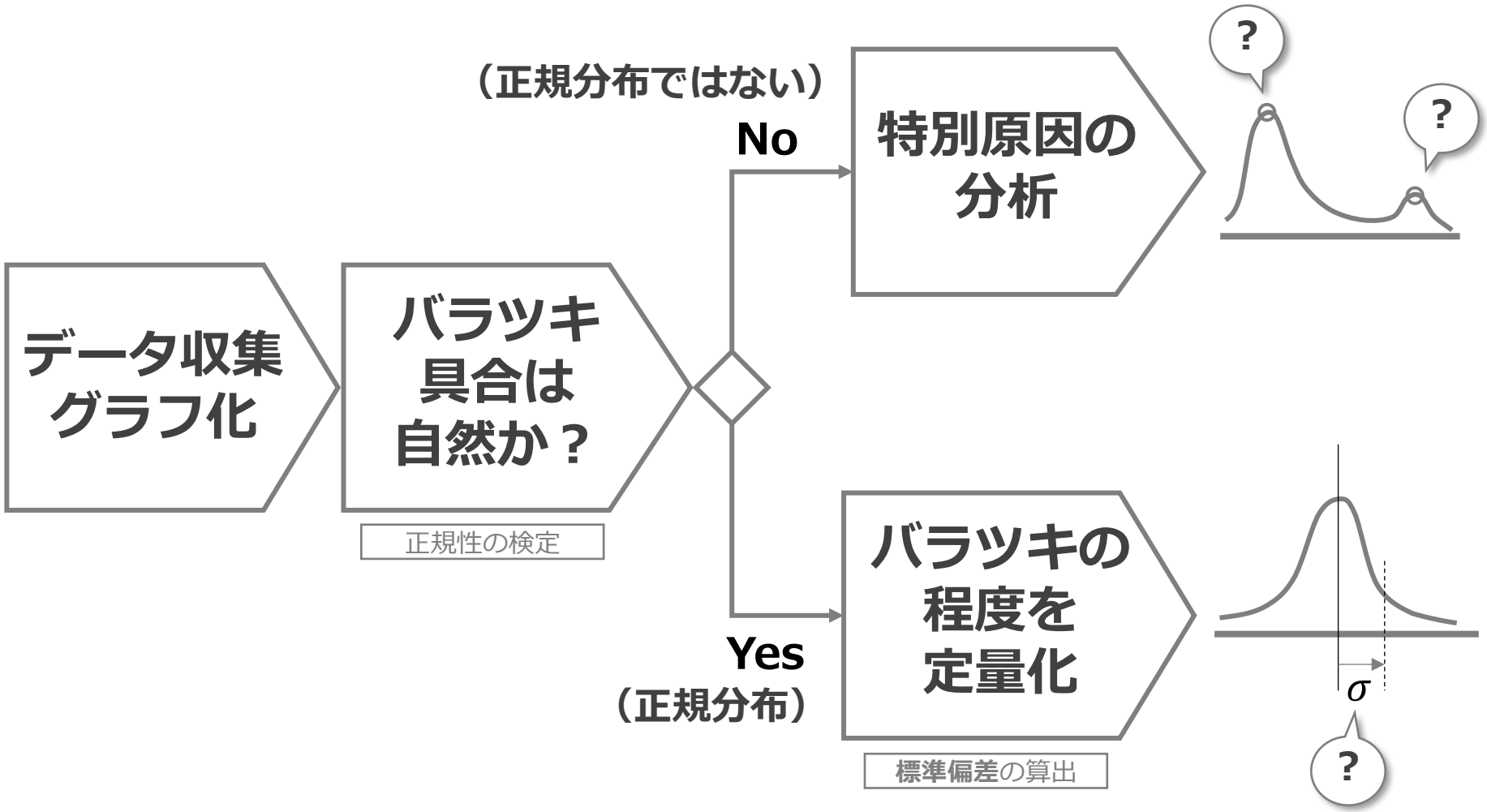
際立った原因が無い時の データの自然なバラツキ具合

- **正規分布** normal distribution：
  - もっとも一般的な分布
  - 連続値のデータが**平均値の付近に集まっている**確率分布
  - **独立な多数の因子**の和の分布
  - 自然科学・社会科学などの**複雑な現象**のモデル
  - **誤差**の大きさの出現確率のモデル



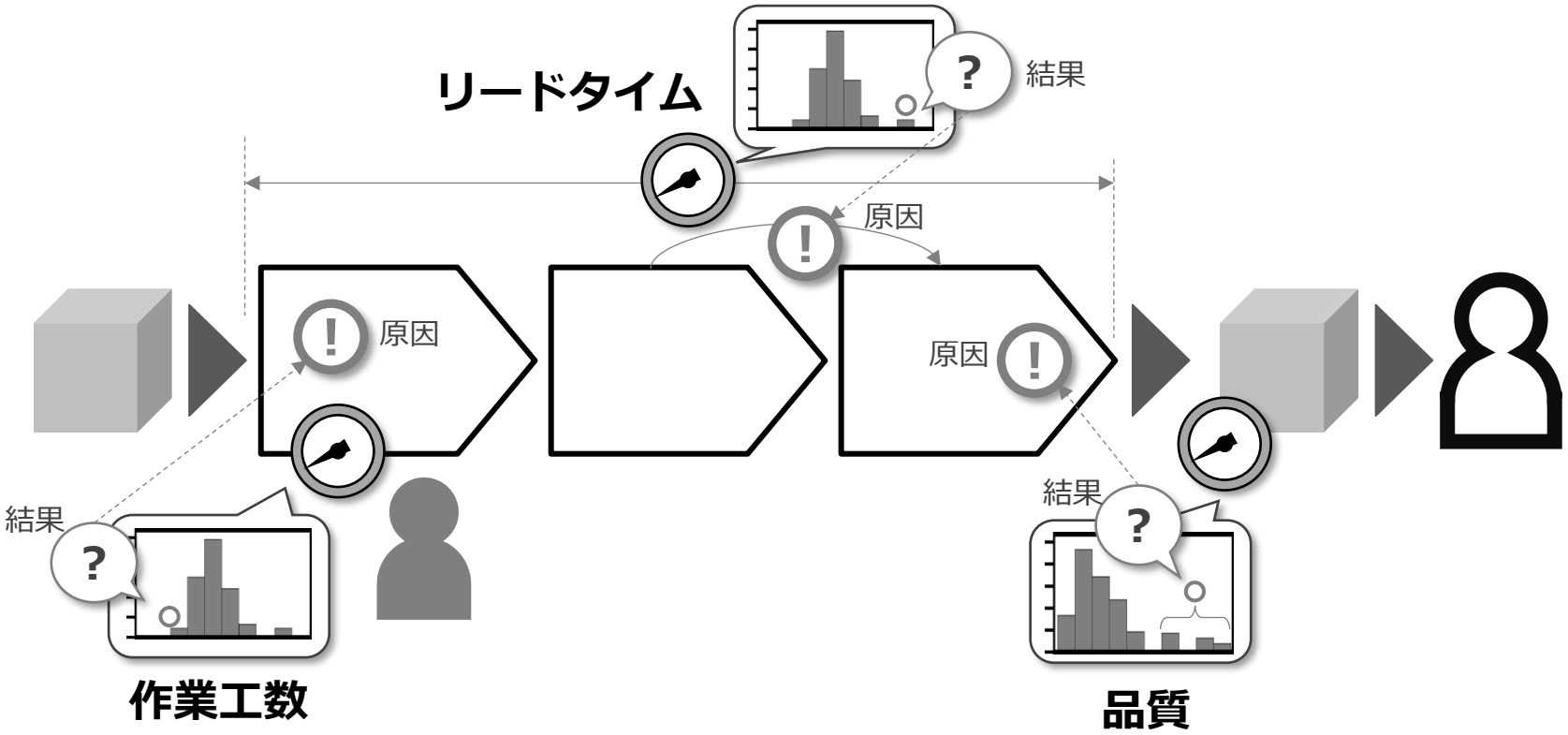
バラツキに着目する

正規分布か否かで 分析や改善の方向性が変わる



バラツキに着目する

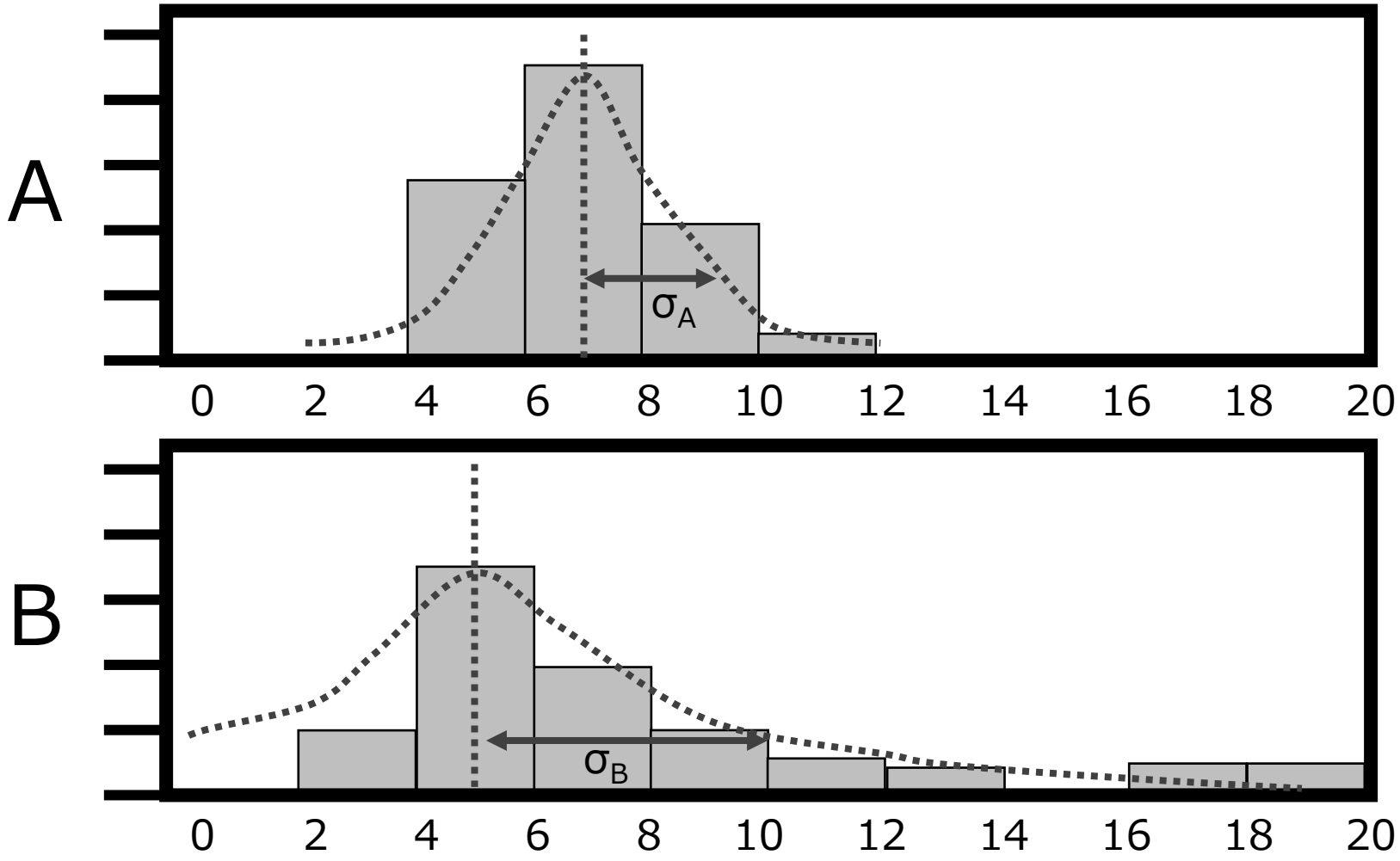
バラツキ具合を見れば、プロセスの問題を見つけやすい



- 平均値だけを眺めていても、問題の原因に迫りづらい
- バラツキが**正規分布でなければ**（偏りがあれば）、大抵の場合プロセスに沿って分析することで、問題の原因を見つけられる

バラツキの定量化

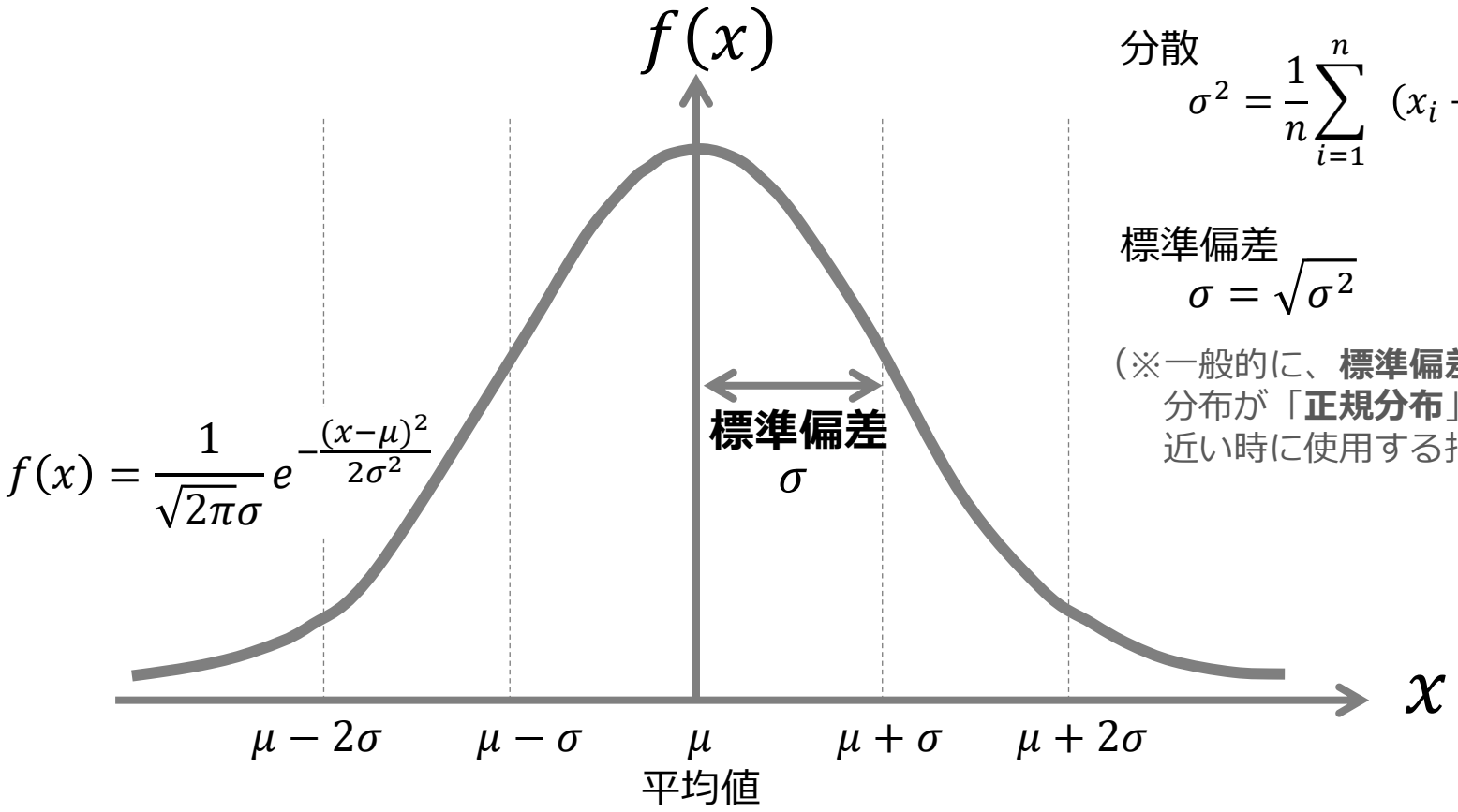
どちらの方が、バラツキが大きい？



バラツキの定量化｜標準偏差

標準偏差は、データのバラツキ具合の指標

- **分散** variance：各データと平均値の差の２乗の和
- **標準偏差** standard deviation ：分散の正の平方根



分散

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

標準偏差

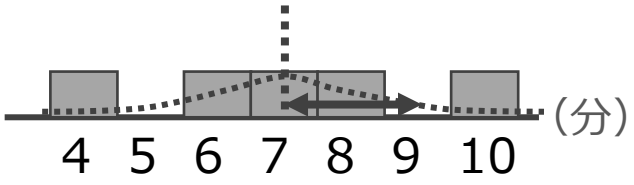
$$\sigma = \sqrt{\sigma^2}$$

（※一般的に、**標準偏差**は、分布が「**正規分布**」に近い時に使用する指標。）

演習｜標準偏差

改善前と改善後で、対応時間のバラツキは変わりましたか

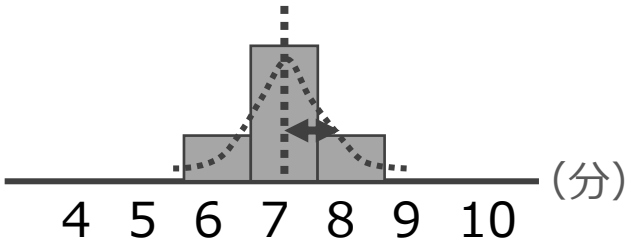
改善前



	時間（分）	(時間-平均)^2
1	4	9
2	6	1
3	7	0
4	8	1
5	10	9
平均	**7**	**4**

分散 ＝ 4
標準偏差 ＝ 2

改善後



	時間（分）	(時間-平均)^2
1	6	1
2	7	0
3	7	0
4	7	0
5	8	1
平均	**7**	**0.4**

分散 ＝ 0.4
標準偏差 ＝ 0.63

まとめ

データ分析

明確な目標定義のもとに
必要十分なデータを集め
適切な観点でグラフ化し
洞察を与え
改善活動に駆り立てよう

参考資料
正規性の検定

仮説検定：帰無仮説と対立仮説

- **帰無仮説** Null hypothesis：
  検証対象の仮説。「差がない」「効果がない」といった仮説を立てて、「現実の観測値から見て、その仮説には無理がある」と棄却する、つまり無に帰す予定の仮説。
- **対立仮説** Alternative hypothesis：
  帰無仮説が成り立たない状態。帰無仮説に無理があると判断した場合、この対立仮説を採択する。
  主張したいことを対立仮説におき、帰無仮説を棄却することで、主張の妥当性を示す。

- **仮説検定** hypothesis testing：
  主張したいことを「対立仮説」におき、
  「帰無仮説」を棄却することで、主張の妥当性を示す。

【注意】帰無仮説を棄却できないとしても、
帰無仮説の正しさが証明されたわけではない。

仮説検定：p値

p値：プロセス改善への適用例

改善前と平均やバラツキが**変わっていないと仮定**した時に
今回のようなデータが測定される**確率**

▼

p < 5% なら、仮定を棄却して **改善効果アリ** と結論

p値：正規性検定への適用例

測定データの母集団が**正規分布であると仮定**した時に
今回のようなデータが測定される**確率**

▼

p < 5% なら、仮定を棄却して **正規分布ではない** と結論

p値とは

- **p値** p-value：
  - 否定したい仮説が、実は、おかしくはない確率
  - 帰無仮説を棄却するリスク（＝過誤を犯す確率）の大きさ
  - 帰無仮説と、現実の観測値が、矛盾しない程度
  - 帰無仮説（＝特定の統計モデル）が正しいと前提した時、実際の観測値およびそれ以上の極端なデータが得られる確率
  - 「差がない」という仮説が前提している統計モデルに従った時、実際の観測値およびそれ以上の「差」が観測される確率
    - 母集団Aと母集団Bの「平均値に差がない」という帰無仮説では、**p値**とは、A,Bからのサンプリングの可能な全ての組合せのうち、「実測した平均値の差」以上の平均値の差が生じる組合せの割合
  - 「改善効果がない」という前提のままでも、実測した改善効果およびそれ以上の効果が生じ得る確率

帰無仮説を棄却する条件は、
p値が有意水準より小さいこと

有意水準とは

- **有意水準** significance level
  p 値の小ささの基準。
  仮説検定において「**第一種の過誤**」を犯す確率（リスク）の許容値。
  - p 値が有意水準よりも**小さい**場合は、
    帰無仮説を棄却するリスクが小さいとみなし、
    帰無仮説を棄却して、対立仮説を採用する。
  - p 値が有意水準よりも**大きい**場合は、
    帰無仮説を棄却するリスクが大きいとみなし、
    帰無仮説を保留する。
- **有意水準**は 0.05 に設定されることが多い。

判断	本当は…	
	帰無仮説が正しい	対立仮説が正しい
帰無仮説を**棄却**	**第一種の過誤** Type I error	正しい
帰無仮説を**棄却しない**	正しい	第二種の過誤 Type II error

p値：コイン投げの例（１）

帰無仮説を**棄却し**、対立仮説を採用するケース：

観測値＝コインを10回投げて表が **1回**しか出なかった
帰無仮説＝表が出る確率は二分の一（普通のコインである）
対立仮説＝表が出る確率は二分の一ではない（イカサマコインである）

p値＝観測値以上に偏った値が得られる計算上の確率
　　＝表が 1回以下か 9回以上出る計算上の確率
有意水準＝0.05
組み合せの数：$2^{10} = 1024$

表の回数	**0**	**1**	2	3	4	5	6	7	8	**9**	**10**
場合の数	**1**	**10**	45	120	210	252	210	120	45	**10**	**1**

$$\textbf{p値} = \frac{1 + 10 + 10 + 1}{1024} \fallingdotseq 0.02 \quad < 0.05$$

帰無仮説を**棄却する**　「普通のコインだとしたら2%の確率でしか起こらない」
　「棄却により第一種の過誤を犯す確率は2%しかない」
→対立仮説を採用する「イカサマコインであると考えることは妥当である」
　（95%以上の確率でイカサマコインと言えたわけではない）

p値：コイン投げの例（2）

帰無仮説を**棄却できない**ケース：

観測値＝コインを10回投げて表が **2回**しか出なかった
帰無仮説＝表が出る確率は二分の一（普通のコインである）
対立仮説＝表が出る確率は二分の一ではない（イカサマコインである）

p値＝観測値以上に偏った値が得られる計算上の確率
　　＝表が 2回以下か 8回以上出る計算上の確率
有意水準＝0.05
組み合せの数：$2^{10} = 1024$

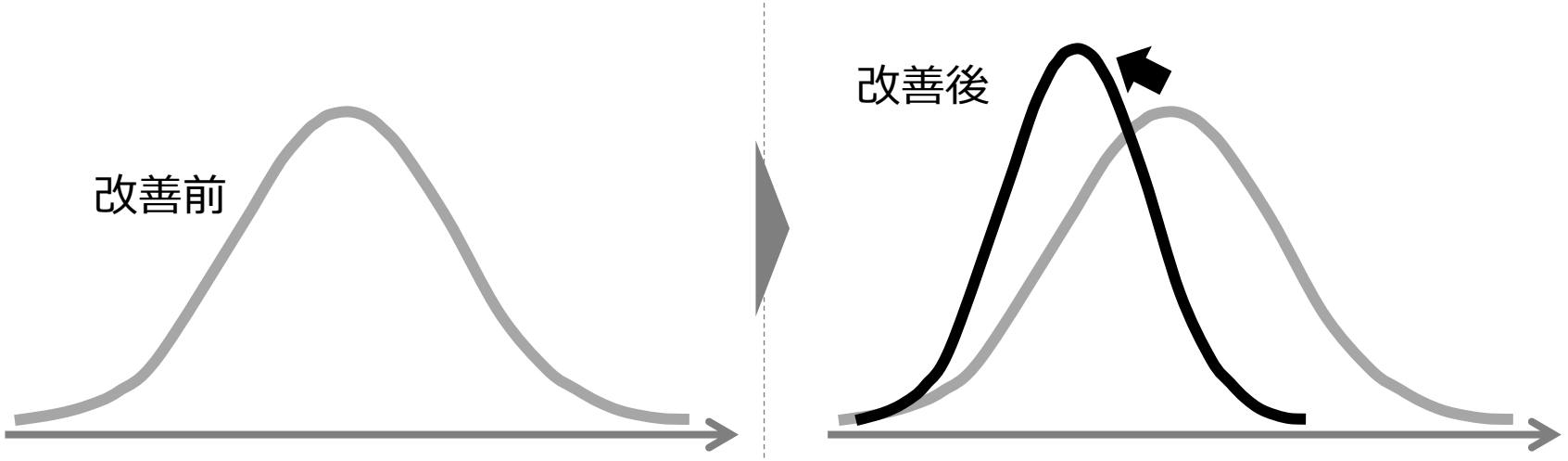
表の回数	**0**	**1**	**2**	3	4	5	6	7	**8**	**9**	**10**
場合の数	**1**	**10**	**45**	120	210	252	210	120	**45**	**10**	**1**

$$\textbf{p値} = \frac{1 + 10 + 45 + 45 + 10 + 1}{1024} \fallingdotseq 0.109 \quad > 0.05$$

帰無仮説は**棄却できない**　「普通のコインだとしても10.9%の確率で起こる」
　　「棄却により第一種の過誤を犯す確率が10.9%もある」
対立仮説を保留する　「イカサマコインであるとは判定できない」
　　（普通のコインであると立証できたわけではない）

正規性の検定

対象が正規分布であると前提できれば、様々な統計技法が使える



改善前・改善後のデータの分布の正規性を検定

- 正規分布**でない**場合
  - 特別要因（欠陥）を見つけてバラツキを抑える
- 正規分布**である**場合
  - **F検定**で、**標準偏差**に有意な改善があったかを確認する
  - **t検定**で、**平均値**に有意な改善があったかを確認する

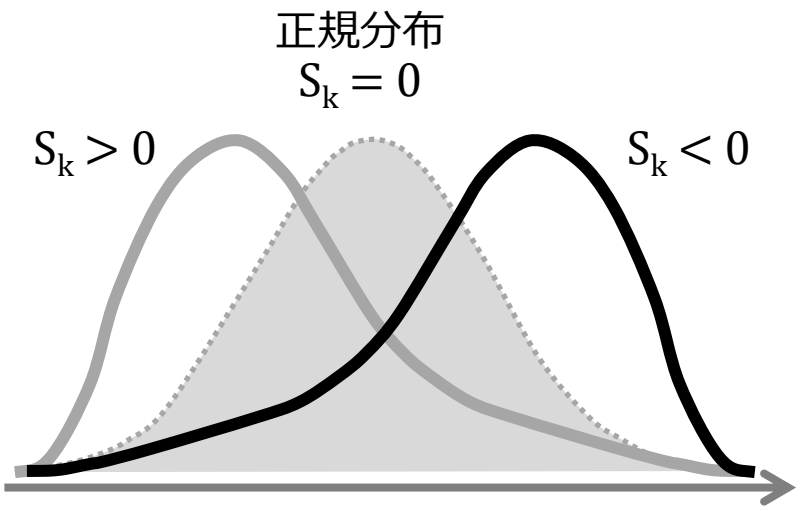
正規性の検定

- **正規性の検定** test of normality：
  データの母集団が正規分布に従っているかを調べる検定。
  - **帰無仮説**：観測値の母集団は正規分布に従っている
  - **対立仮説**：観測値の母集団は正規分布に従っていない

- p値 ≦ 0.05であれば
  「観測対象が正規分布に従っていると仮定して、
  　実際の観測値の分布が得られる確率は、
  　5% よりも小さい」（＝第一種の過誤を犯すリスクは小さい）
  ので、安心して帰無仮説を棄却し、
  「正規分布ではない」という対立仮説を採用する。
- p値 ＞ 0.05であれば、
  「観測対象が正規分布に従っていると仮定し、
  　実際の観察値の分布が得られる確率は、
  　5% よりも大きい」（＝第一種の過誤を犯すリスクを見過ごせない）
  ので、正規分布に従っていないと言うことはできず、
  正規分布であるという仮説を保留する。

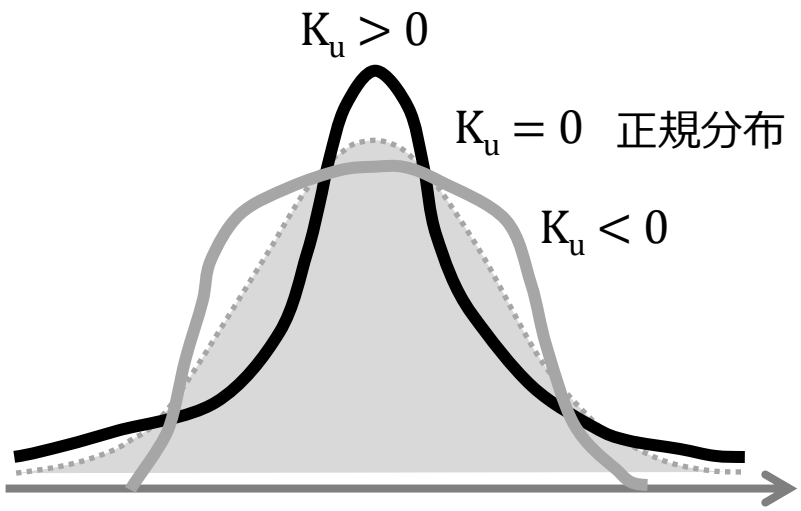
正規性の検定

正規分布との差は、「歪度」と「尖度」で測る



歪度
skewness

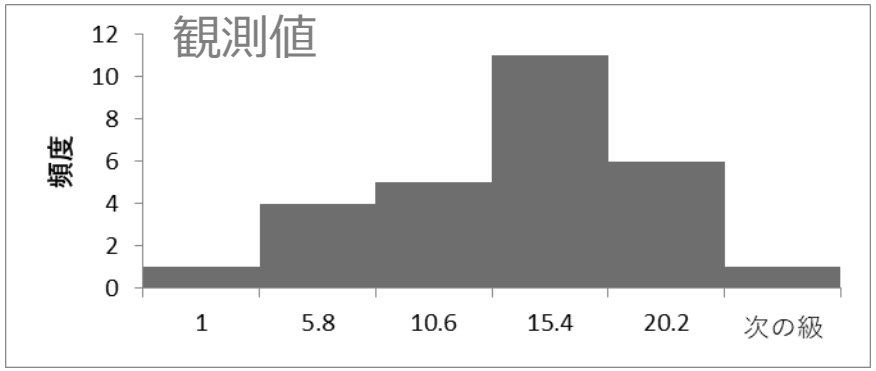
分布の非対称性



尖度
kurtosis

分布の鋭さ

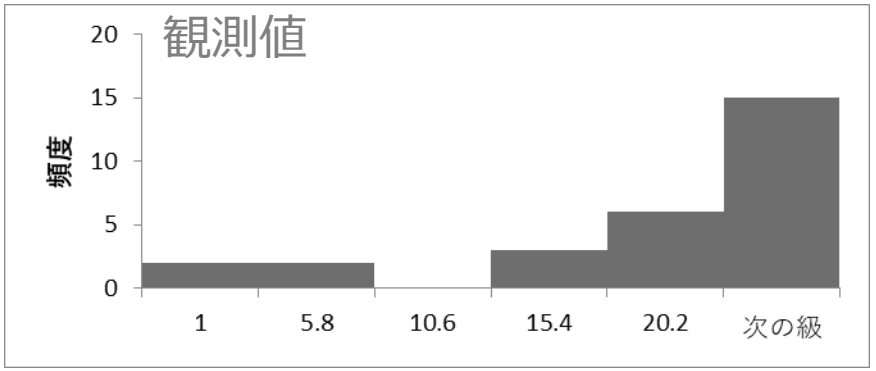
正規性の検定



平均値	11.714
標準偏差	5.798
歪度	0.003
尖度	-0.203
p値(※)	0.893

正規分布の可能性がある

※正規分布である という
帰無仮説を棄却できない



平均値	18.357
標準偏差	8.075
歪度	-1.164
尖度	0.157
p値(※)	0.041

正規分布ではない

※ p値 = D’Agostino and Pearson検定（K2検定）

正規性検定でのｐ値

正規性検定での
「ｐ値」の意味が
良く分からない…

「ｐ値」が5%より**小さければ**、
正規分布じゃなくて、**<u>特別な要因</u>**がある
と考えよう。

（※「本当は正規分布なのに、そうではないと
誤って判断するリスク」が5%より**小さい**
→「本当は正規分布」という帰無仮説を棄却する。）

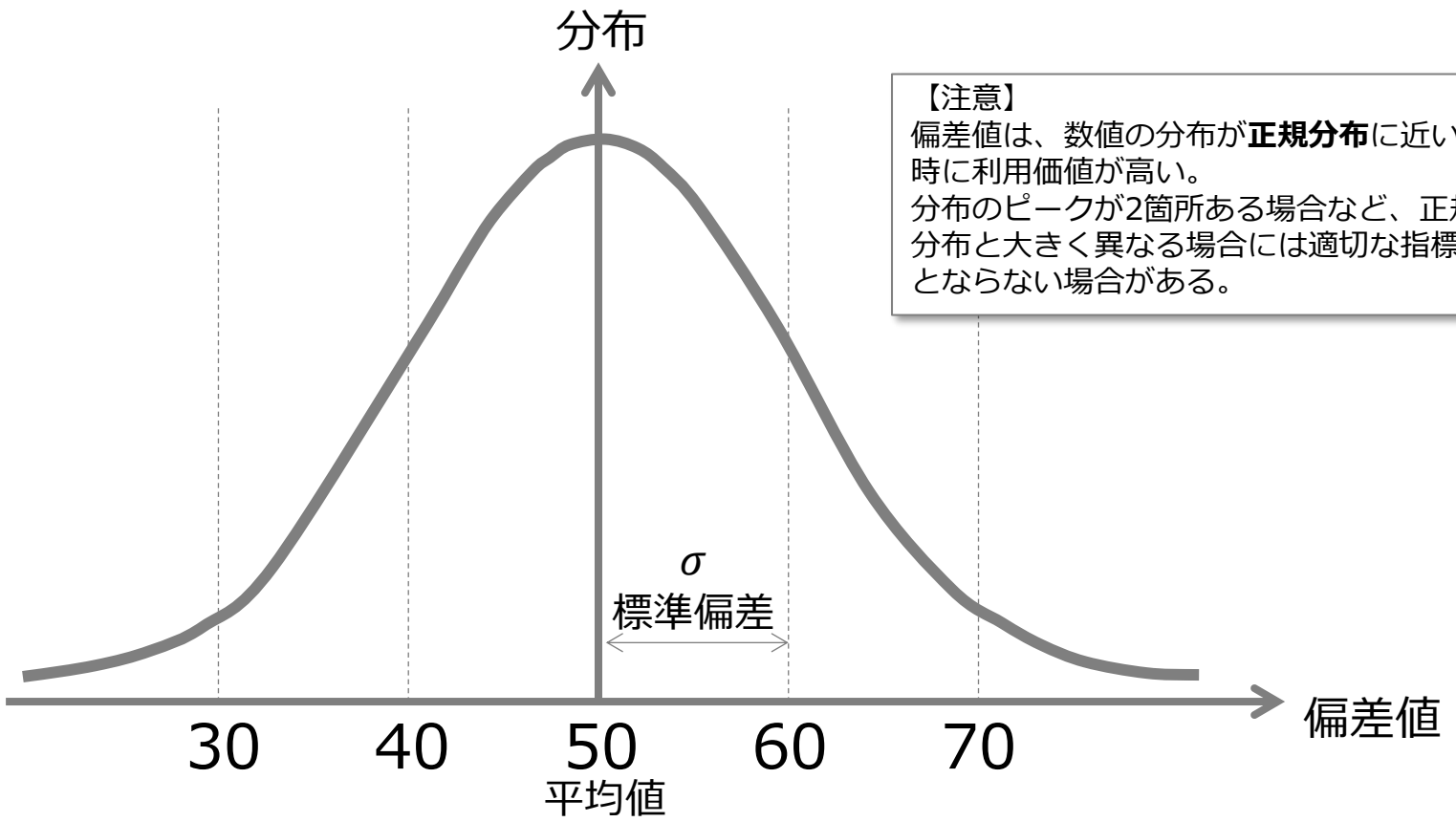
「ｐ値」が5%より**大きければ**、
正規分布でないとは言い切れない、
特別な要因があると言い切れない、
ってことね！

（※「本当は正規分布なのに、そうではないと
誤った判断をするリスク」が5%より**大きい**
→「本当は正規分布」という帰無仮説を棄却できない。）

偏差値

偏差値とは、ある数値が全体の中のどの位置にいるかを表した数。

- **偏差値** standard score：平均値が50、標準偏差が10となるように標本変数を規格化したもの



【注意】
偏差値は、数値の分布が**正規分布**に近い時に利用価値が高い。
分布のピークが2箇所ある場合など、正規分布と大きく異なる場合には適切な指標とならない場合がある。

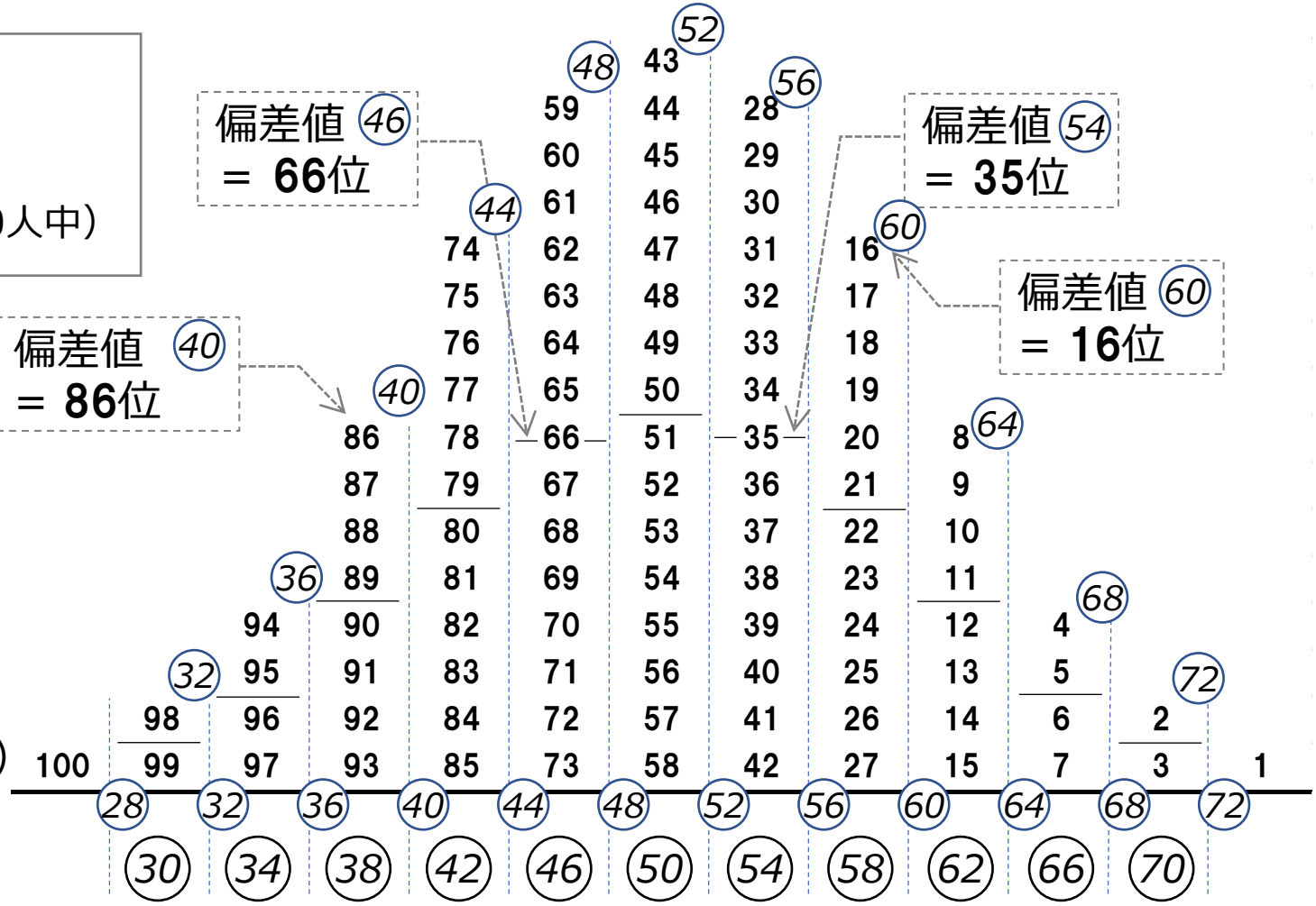
偏差値と順位の目安

（正規分布の場合）

【凡例】

(50)：偏差値

50：順位（100人中）



プロセス能力 process capability index

製品の品質や特性を安定して作り出せる能力

$$\mathbf{Cp} = \frac{\text{USL（上方規格値） - LSL（下方規格値）}}{6 \times \text{標準偏差}\sigma}$$

（※1.33 < cp < 1.67 で十分な工程能力とされ、
不良が発生する確率は　0.6ppm～63ppm 程度）

データの平均値（μ）が規格範囲の中心値（CL）とは限らないため、
そのズレを考慮した工程能力 **Cpk** も良く使われる

計算式①　$\mathbf{Cpk} = \min\left(\frac{USL - \mu}{3 \times \sigma} , \frac{\mu - LSL}{3 \times \sigma} \right)$

計算式②　$\mathbf{Cpk} = \left(1 - \frac{|（USL + LSL）/ 2 - \mu|}{（USL - LSL）/ 2} \right) \times Cp$

分子：規格中心値と
平均値の差の絶対値

分母：規格幅の半分

参考資料

数の起源

- 原始人は、２や３以上の数を表わす言葉や記号を持たなかった
- 人間は進化し、抽象的な†『数』を扱えるように進化
- 数を表わす記号（数字）も、文明と共に発達

- 人類は有史以来、抽象的な「数」の概念を理解するために途方もない時間をかけてきた。
- 「数」は人類の**至宝**・素晴らしい**共有財産**

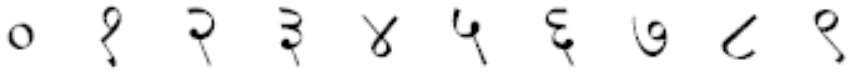
† 『数』は抽象的である。実数は「実」と書かれるし、自然数は「自然」と書かれる。しかし、「１個」の石に つまづいて転んだ人はいても、「１」という『数』に つまづいて転んだことのある人はいない。虚数は「虚」と書かれるが、実数も虚数も具体物ではないという意味では同じである。

数字の歴史

■**エジプト文明**（紀元前3,000年頃）

■**メソポタミア文明**（紀元前3,000年頃）

■**インド**（紀元前2,500年頃）（デーバナーガリー文字）



■**ギリシア**（紀元前800年頃）

■**ローマ**（紀元前300年頃）

I II III IV V VI VII VIII IX X C D M

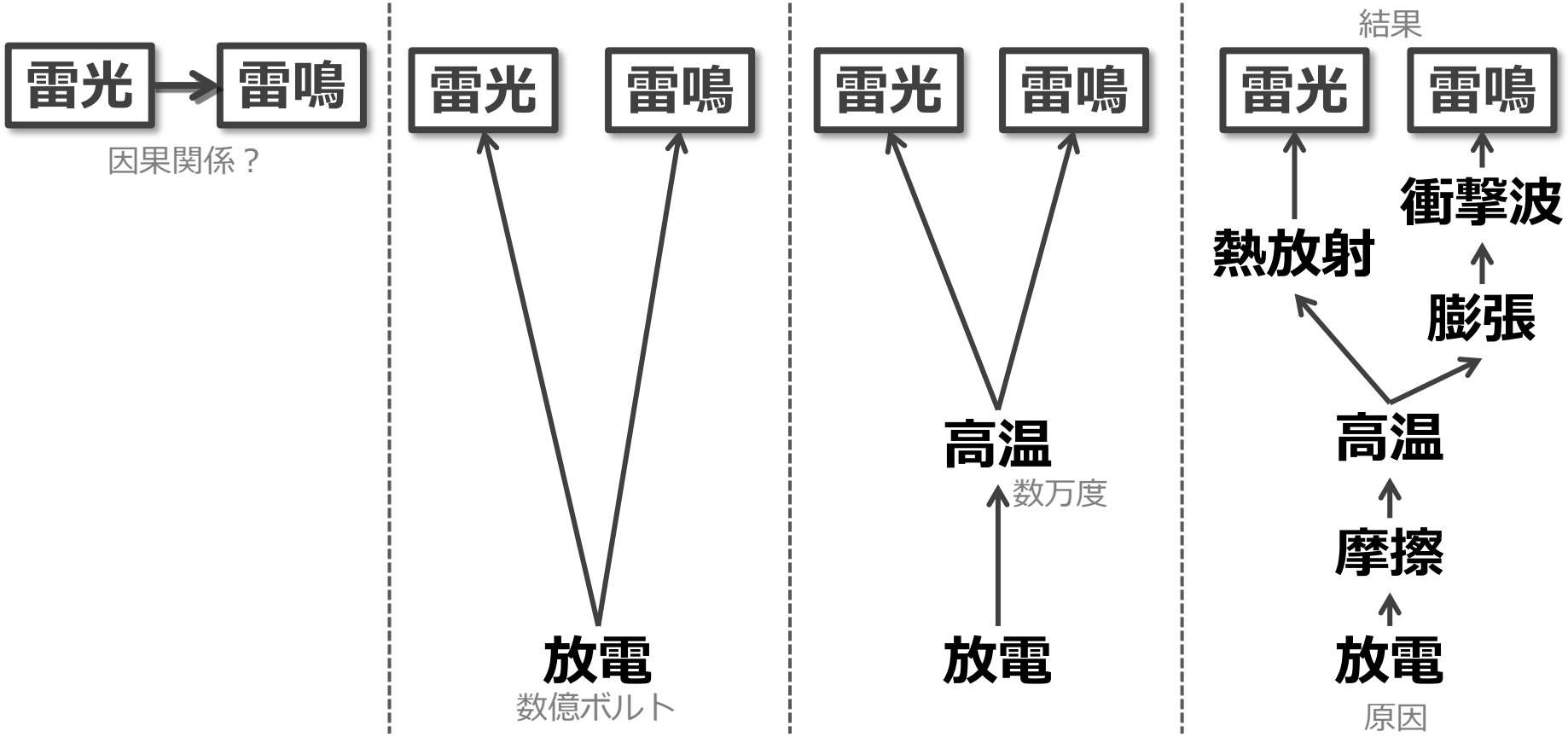
十 百 千

ゼロの発明

- 数は「ものを数える言葉」で、１ から始まっていた
- 「物がある」という概念を量的に表したのが数であり、「物がない」という概念には「無い」という言葉が使われた
- 数字を並べる「位取り」（１０進法など）が現れると「数のない桁」を表わす記号が出現した
  - 紀元前５００年頃のくさび型文字に位が０である記号が表れ始めた
  - １世紀頃のインド算盤で、数のない桁を書き写す記号が発明された
- 628年インド人数学者ブラーマグプタが「数」としての「0」の概念を見出し、現代に近い計算法を案出した

因果関係 と 相関関係

ピカッと光ったから、ゴロゴロ鳴るの？



雷光と雷鳴は、実は **因果関係**ではなく、**相関関係**

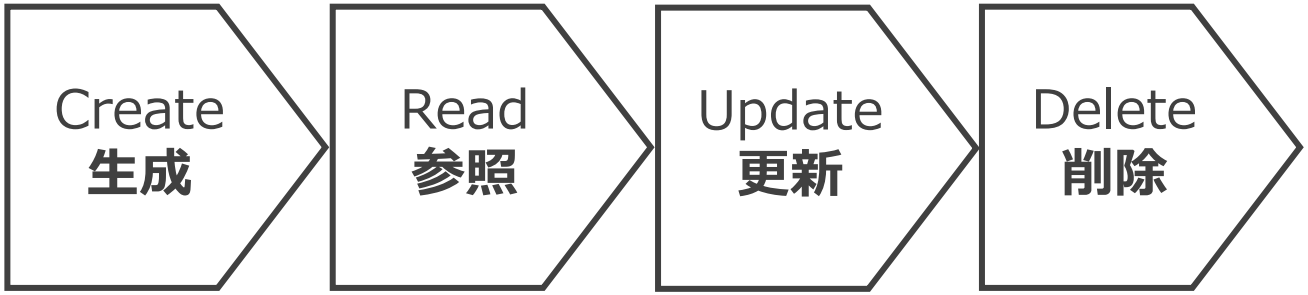
現実世界の因果関係は、経験則・信念に過ぎず、究極的には証明できない

因果関係 と 相関関係

- 究極的には、因果関係は証明できない
- Ａになると、引き続いてＢになることが、
  経験的に確からしい時、「Ａ＝原因、Ｂ＝結果」
  という**因果関係**（A→B）と見なされる
  - 別の原因Cから C→A、C→B が生じている、と
    分かれば、Ａ と Ｂ は**相関関係**であると判明する
- 因果関係は、原理的・絶対的な関係ではなく、**経験**や**信念**。
  - 物理法則も、経験と信念が積み重なったもの。
  - 因果関係は、**創る**こともできる。
    - 例えば、ビジネスとは、商品（製品やサービス）を原因、
      お金を結果とする、因果関係を創ること。
      （生産・営業などのプロセスを仕組み化すること。）
  - A→B間のプロセスは、詳細に可視化すれば
    より複雑な因果関係の連鎖に分解できるし、
    それゆえ常に**改善**の余地を持っている。

CRUD（クラッド）サイクル

ソフトウェアが満たすべき、データ処理機能の種類



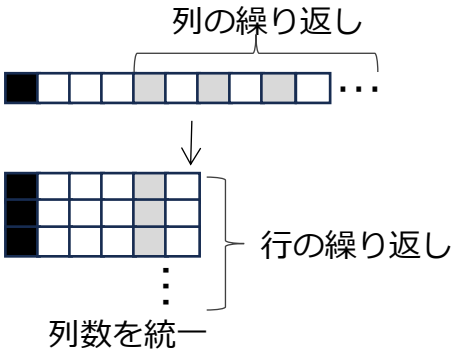
- あらゆる情報の扱いにあてはまる
- 機能・プロセス・ルールの検討漏れを防げる

※ データベース管理システムが備えるべき主要機能が由来

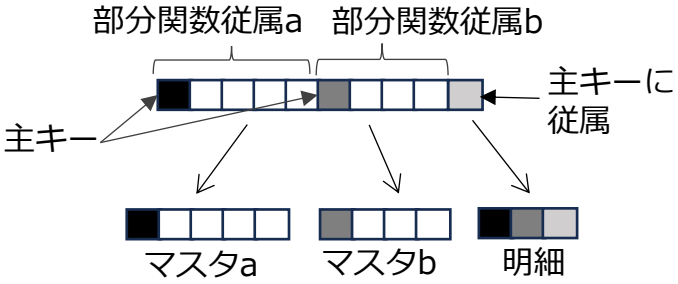
データの正規化

正規化によってデータの保守性が高まる

第一正規形

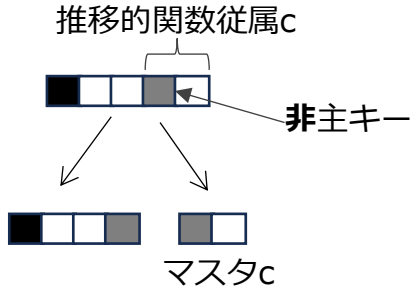


第二正規形



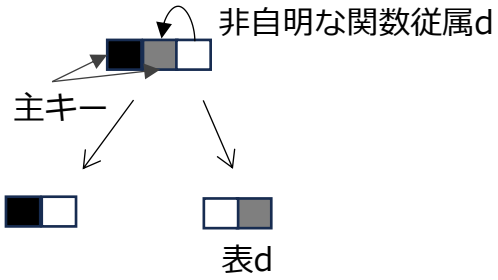
主キーについて
重複を分離

第三正規形



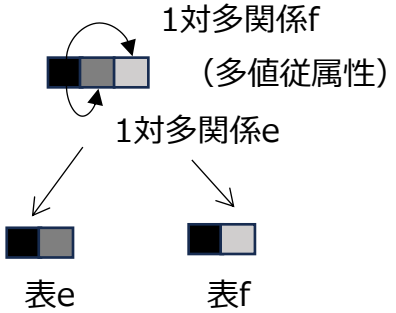
非主キーについても
重複を分離

ボイスコッド正規形



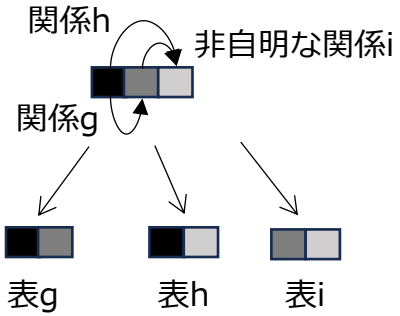
隠れた従属の分離

第四正規形



多対多の重複が無い
（更新の冗長性を排除）

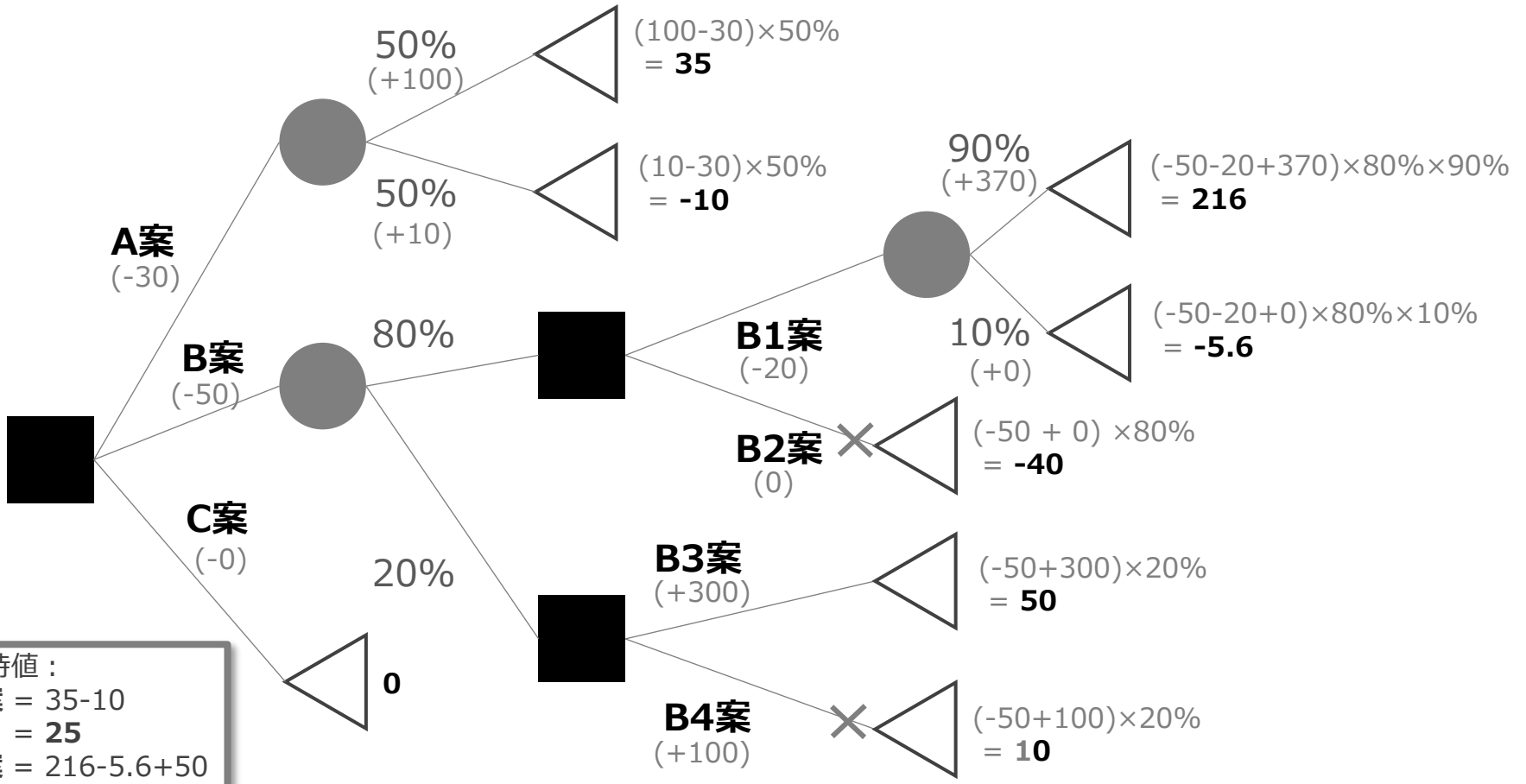
第五正規形



分解不可能な最小単位
（全ての関係を形式化）

ディシジョンツリー

どの選択が最も正しそうかを定量的に決定する



期待値：
A案 = 35-10
= **25**
B案 = 216-5.6+50
= **260.4**
C案 = **0**

KJ法

定性データをグループ化し分析する汎用的な手法

コロナ禍における職場の課題						
インフラ・設備	マインド・メンタル・健康	風潮・風土	マネジメント	仕事の進め方	コミュニケーション	会議
3	4	2	4	5	1	5
ITインフラが不十分	会話減ってウツ気味	コロナ禍を機会と捉える風潮がない	部下在宅勤務状況を上司が監視したがる	現場に行けないので仕事のやり方が変化	飲み会が減った	17時以降の会議は家族に迷惑がかかる
VPN接続の性能が不十分	ストレスがたまる	出勤すると嫌がられる	細かい作業が確認できない	紙での印刷の方がチェックしやすいので、困る		遠隔会議では相手に伝わったか分かりづらい
自宅に大きなモニターがなく視力低下	ミーティングが減って孤独を感じる		メールでの部下との会話は膨大な時間が必要	仕事の組み立て方に毎回悩む		遠隔会議は対面会議よりも疲れる
	運動が減って体重増加		評価しづらい	在宅業務で出来ないことを省略してしまう		移動時間なくビッシリ遠隔会議が入る
				直接会話できないので細かい確認が抜ける		うっかり自宅の状況が見えると恥ずかしい

すべてはお客様の
「わかった」
「なるほど」
「やってみよう」
のために

Current Color