情報処理応用B 第11回

藤田 一寿

問題解決と統計処理

分析と改善

- 問題が起こったときにどう解決すればよいのか.
- 現状を分析し改善する.
- 現状分析の手法
  - 統計処理
- 改善の手法
  - PDCAサイクル

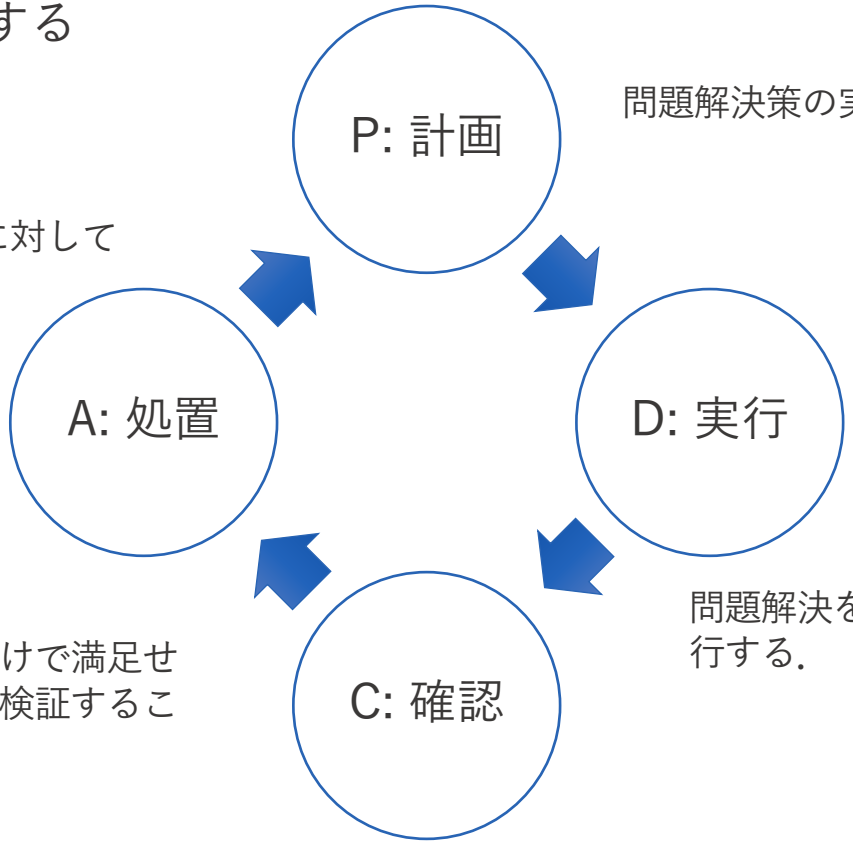
PDCAサイクル

PDCAサイクル

- 改善のための4つの手順(PDCAサイクル)
  - ① 計画(P: Plan)
    - 目的を決め、達成に必要な計画を設定
  - ② 実行(D: Do)
    - 計画に基づき実施
  - ③ 確認(C: Check)
    - 実施の結果を調べ評価
  - ④ 処置(A: Action) 処置
    - 必要により適切な処置

PDCAサイクル

PDCAサイクルで解決する
テーマを設定する



問題解決策の実施を計画

問題解決を計画したものを実行する.

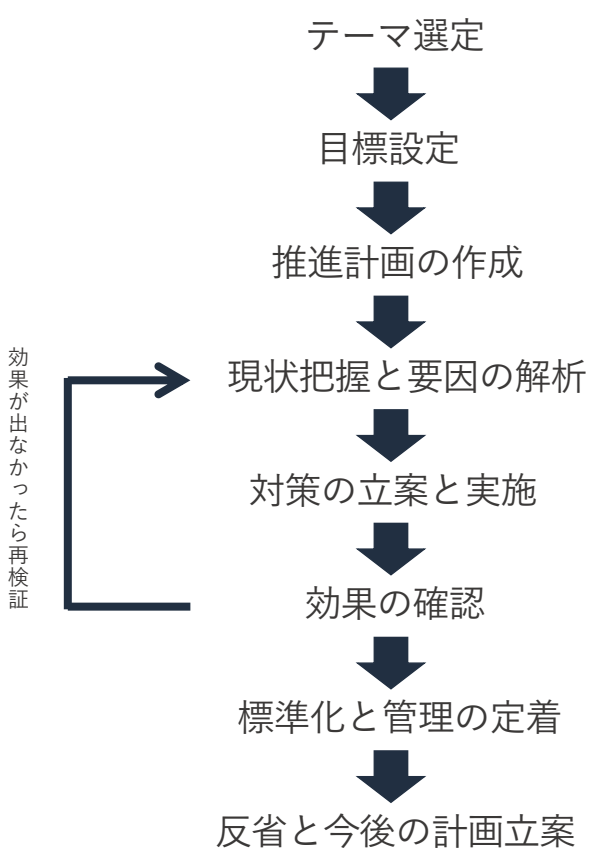
実行結果を解析し問題を見つける.

実行するだけで満足せず、それを検証することが重要

見つけた問題結果に対して解決策を立てる.

確認をして目的を達成していればPCDAサイクルを抜けて良い.

改善のアプローチ



問題点を洗い出す。重要で解決の可能性があるものを考慮する。

目標はできるだけ数値化する。目標が大きい場合は第1次、第2次と分割する。

協力体制を作る。分担を決める。

テーマについて原因を詳しく調査する。調査結果を分析し、問題に対する真の原因を見つける

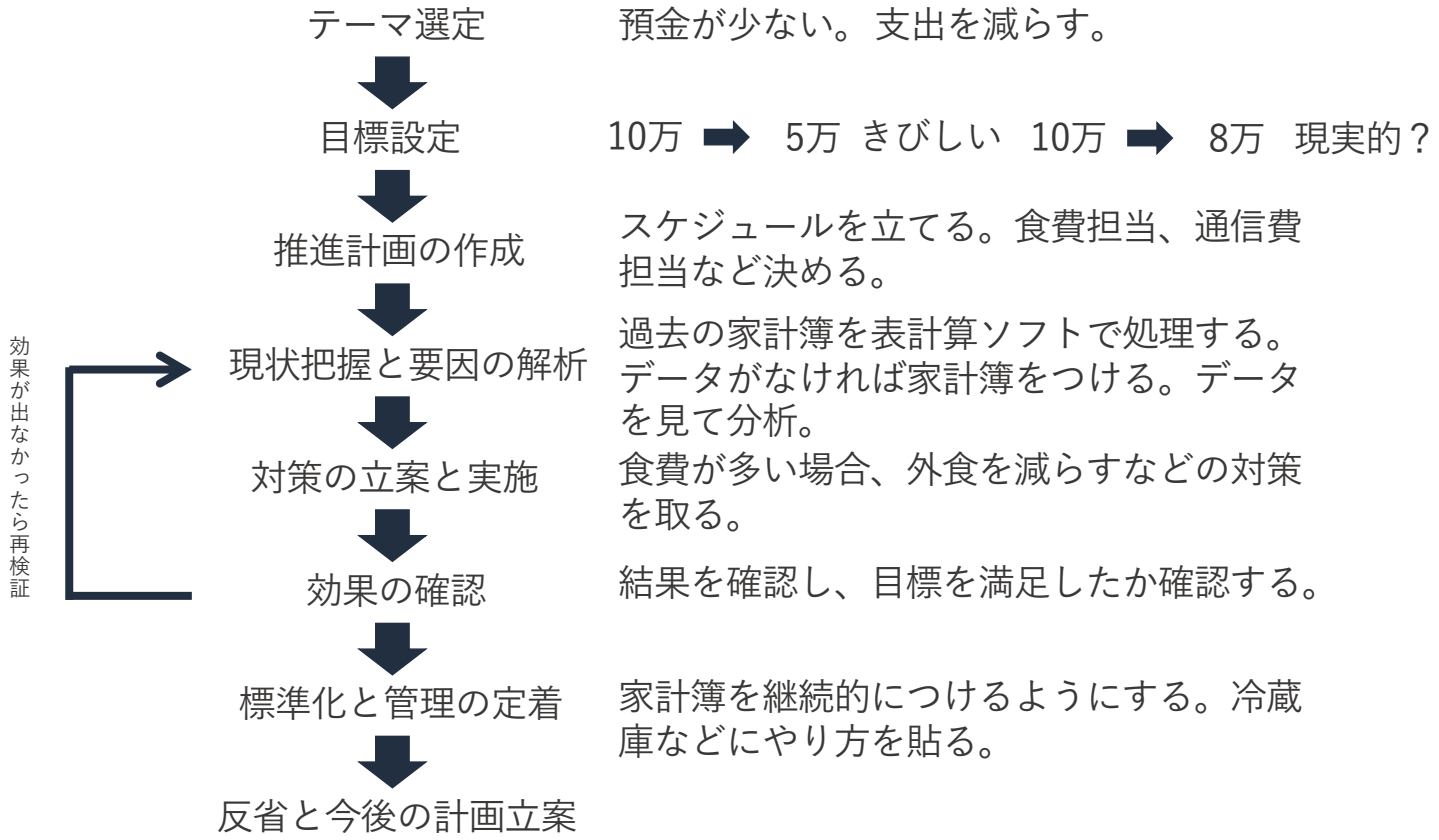
把握した原因に対し有効な対策を考え実施する。

結果を確認し、目標を満足したか確認する。

標準の制定、改定や治工具類の作成などにより再発の防止を図る

問題解決方法の共有が重要

家計支出の例


テーマ選定
預金が少ない。支出を減らす。
目標設定
10万 ➡ 5万 きびしい 10万 ➡ 8万 現実的？
推進計画の作成
スケジュールを立てる。食費担当、通信費担当など決める。
現状把握と要因の解析
過去の家計簿を表計算ソフトで処理する。データがなければ家計簿をつける。データを見て分析。
対策の立案と実施
食費が多い場合、外食を減らすなどの対策を取る。
効果の確認
結果を確認し、目標を満足したか確認する。
標準化と管理の定着
家計簿を継続的につけるようにする。冷蔵庫などにやり方を貼る。
反省と今後の計画立案
効果が出なかったら再検証

PDCAサイクルは使えるか

- PDCAサイクルは使えないと言われることがあるが，まだ有用な手法である．

- テーマ（目的）設定を適切にすることが重要である．
  - 目的が間違っていれば，サイクルを回しても意味がない．
- テーマ（目的）をサイクルの途中で変更しない．
  - 目的がコロコロ変わると何をすればよいか分からなくなる．
- テーマ（目的）を達成したらサイクルを抜けてよい．
  - 目的を達成したにも関わらずサイクルを回す意味がない．
- 問題を解決した方法を共有する．

- PDCAサイクルは何のために回すか必ず心に留めておく．
- 逆に何のために回しているかわからなければPDCAサイクルは役に立たない．

統計の基礎

科学的アプローチ

- 問題を解決するには，経験や感だけではなく，データや理論を用いて現状を認識する必要がある.

- 数値化する→定量化により客観的に結果を評価する.
- 可視化してわかりやすく→グラフ

- 数値化と可視化の両方が重要

問題解決で統計やグラフが必要なわけ

- 現状を数値化し評価したい
- すべての製品をチェックできないが製造工程全体を評価したい
- 現状を可視化してわかりやすく理解したい
- 図で表すことで問題点を発見しやすくしたい
- など

なぜ統計が必要か

- 無数にあるデータ一つ一つをチェックできない
- 個々のデータをみるだけでは全体の傾向がつかめない

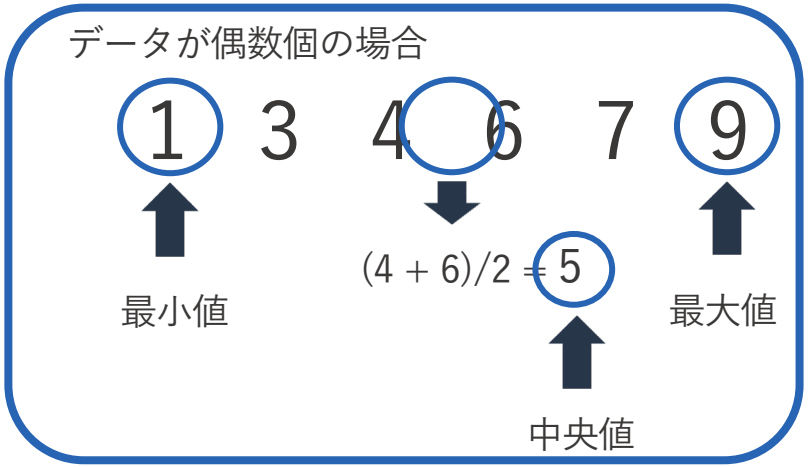
統計的手法を用いる

統計でよく用いる基礎的な数値（統計量）

- データ数
- 最大、最小
- 中央値（順番的に真ん中）
- 平均（データ重心）
- 分散、標準偏差（ばらつき具合）

最大値、最小値、中央値

- 最大値
  - 最も大きい値
- 最小値
  - 最も小さい値
- 中央値
  - 順番的に中央の値
  - 平均より中央値の方が適切な場合もある。平均は外れ値に引っ張られるため。



平均（算術平均）

- N個のデータ $x_1, x_2, \ldots\ldots, x_N$ の平均は下記のように表される.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

- データの重心の意味も持つ.
- 平均は外れ値に引っ張られるため，データを表す数値として不適切な場合もある.

平均と中央値

- 平均と中央値どちらが集団の特徴をより表しているのか？

- 右の例
  - 平均点：6.85点
  - 中央値：8点

  - 平均点は0点と1点の人に引っ張られていて，中央値にくらべ低めの値になっている．

- 統計量の特徴を知っておかないと，状況の把握を間違えることがあることに注意する．

あるテストの得点と人数

点数	人数
10	2
9	4
8	5
7	4
6	2
5	0
4	0
3	0
2	0
1	1
0	2

分散、標準偏差

- 分布のばらつき具合を表す指標
- 分散

$$V = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x})^2$$

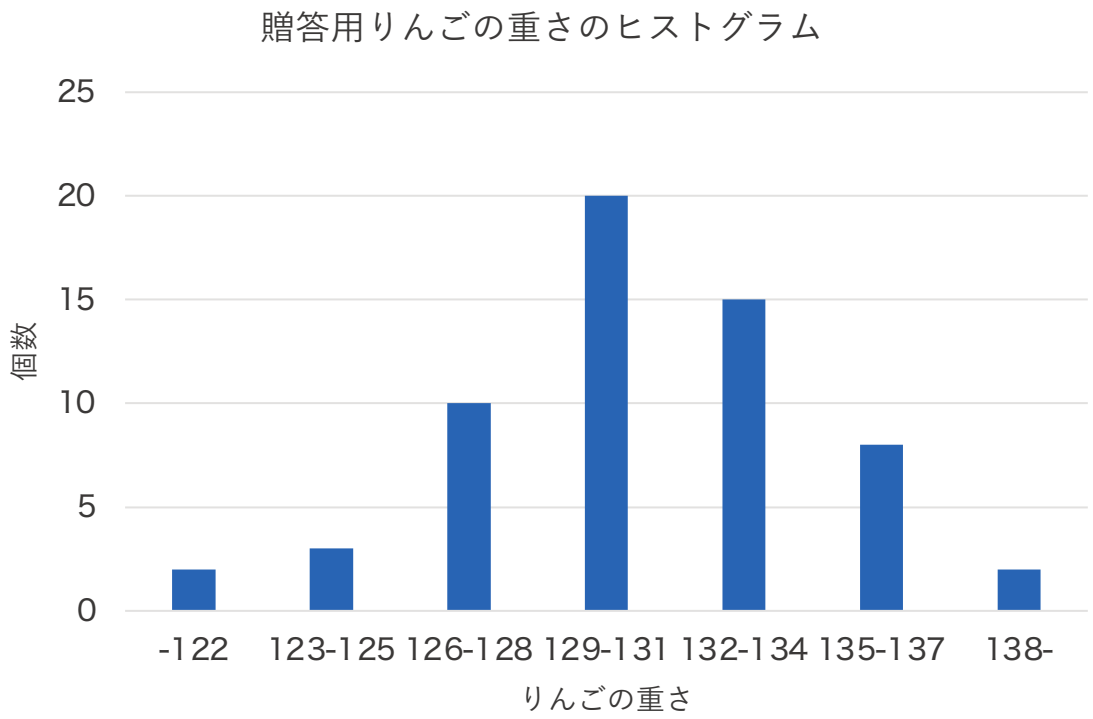
  - 各データの平均からの距離の2乗の平均
- 標準偏差

$$S = \sqrt{V}$$

$$(\sigma)$$

ヒストグラム(度数分布)

- データの分布の様子の把握に用いられる.



ヒストグラムの作り方

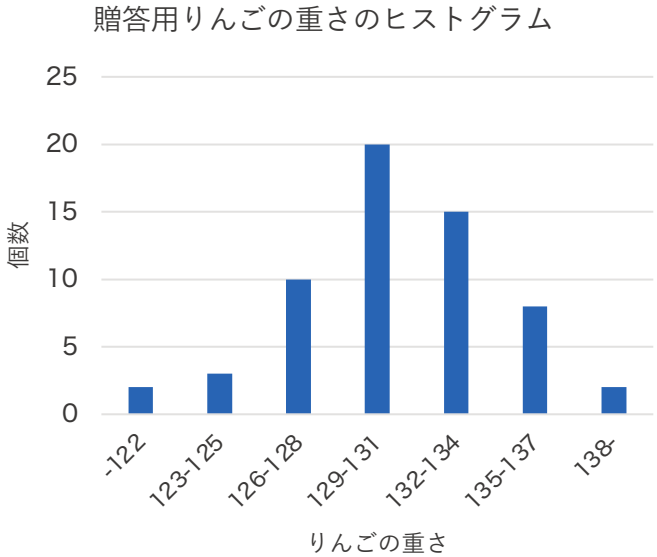
- 度数分布表を作る.
  - 観測値がとりうる値をいくつかの階級に分ける.
  - 観測値がそれぞれの階級でいくつあるか数える.
- 度数分布表にもとづき棒グラフを書く.
  - この棒グラフをヒストグラムという.

リンゴの重さのデータ表(g)

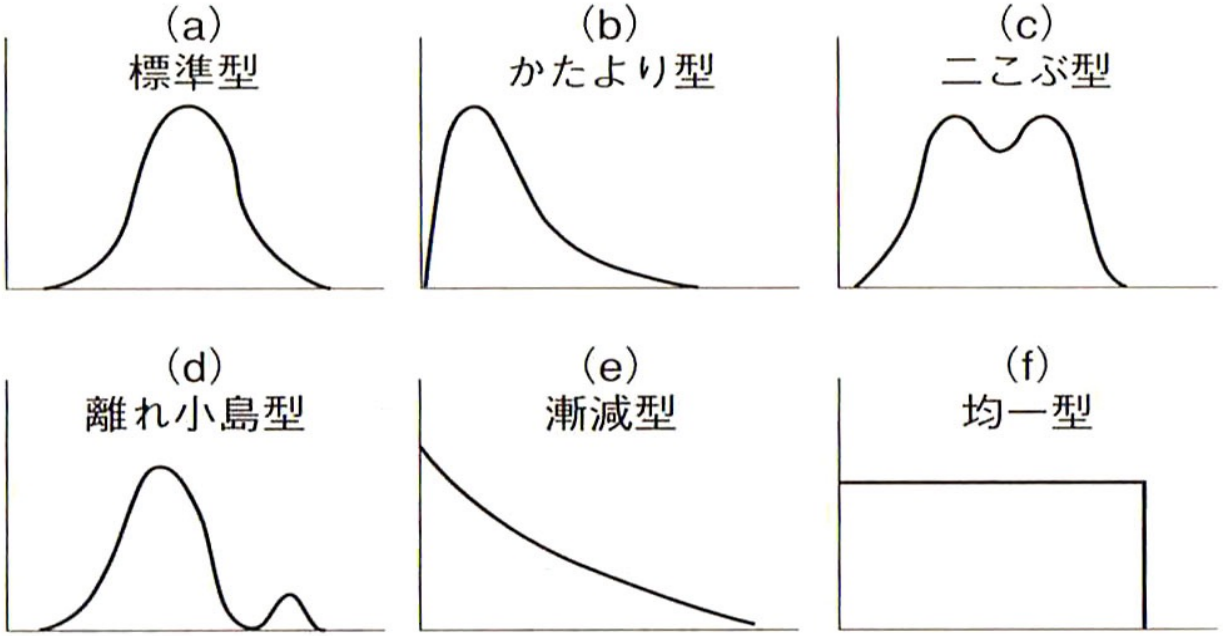
133	130	127	121	137	130
132	130	129	130	130	137
135	133	121	129	132	130
140	133	132	129	129	132
126	132	132	127	129	129
130	124	135	137	127	132
126	129	130	135	137	132
130	130	127	133	135	124
126	127	130	132	133	126
124	127	140	130	132	129

度数分布表

階級	度数
-122	2
123-125	3
126-128	10
129-131	20
132-134	15
135-137	8
138-	2



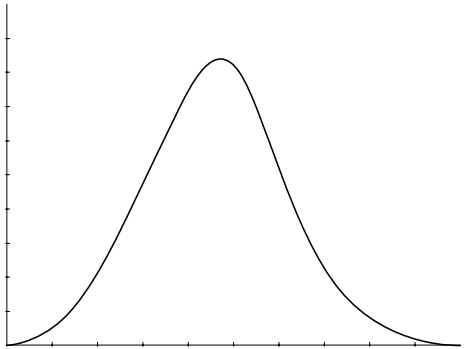
分布の形状



（QC数学の話 大村平より）

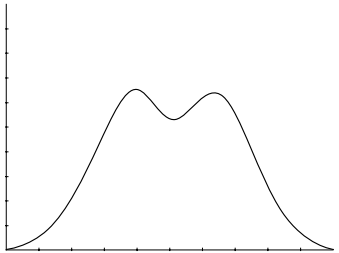
正規分布

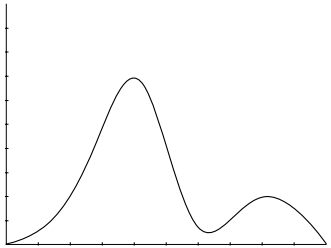
- 最も基礎的な分布の形
- 正規分布と呼ぶ
  - 標準型、一般型、ベル型、ガウス分布などと呼ばれることもある
- 何かを測定した場合、この分布になることが多い。



ふたこぶ型

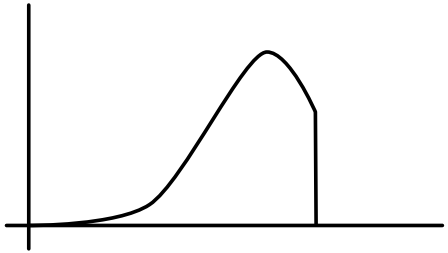
- 複数の要素を含む場合に生じる.
  - 成績の分布
    - 理解している人の集団と理解していない集団がある.
  - 製品の形状の分布
    - 一部が違う規格で作られている可能性がある.
    - このような分布が見られた場合、その原因を探る必要がある.
- データを適切にグループ分けすることで，峰が一つの単純な分布になることが多い．このグループ分けを層別と呼ぶ.



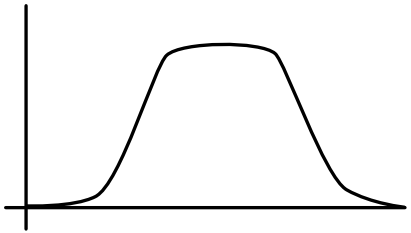


その他

- 絶壁型
  - ある値以下もしくは以上のものを選別して取り除いたときに現れる分布。

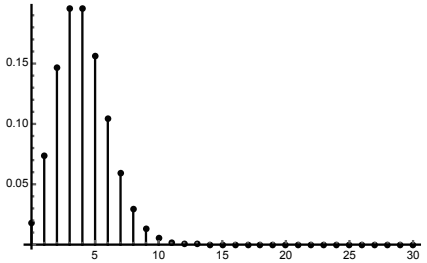


- 高原型
  - ふたこぶ型の一種
  - 平均値が少し異なるいくつかの分布が混在したときに現れる分布。
  - 層別して原因を探る必要あり。

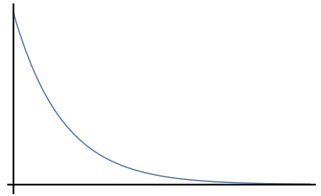


その他

- ポアソン分布
  - 品質管理の世界では偏り型などと呼ばれることもある。
  - 交通事故件数，大量生産の不良品件数、火災件数などはこの分布になる.

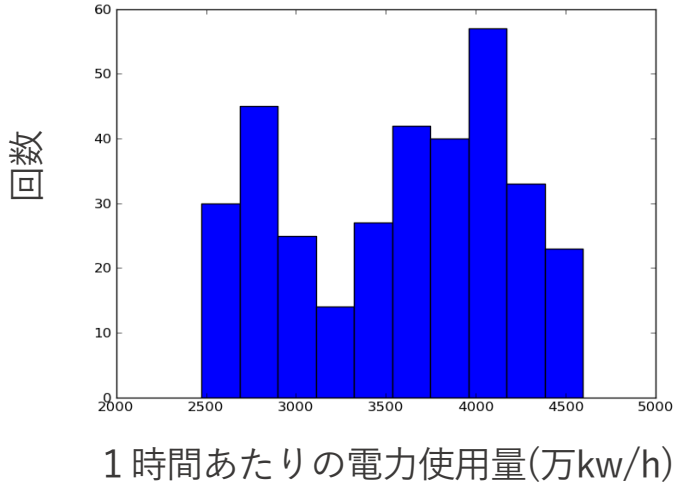


- 指数分布
  - 品質管理の世界では漸減型などと呼ばれることがある.
  - 待ち時間，製品の故障、寿命などはこの分布になる.



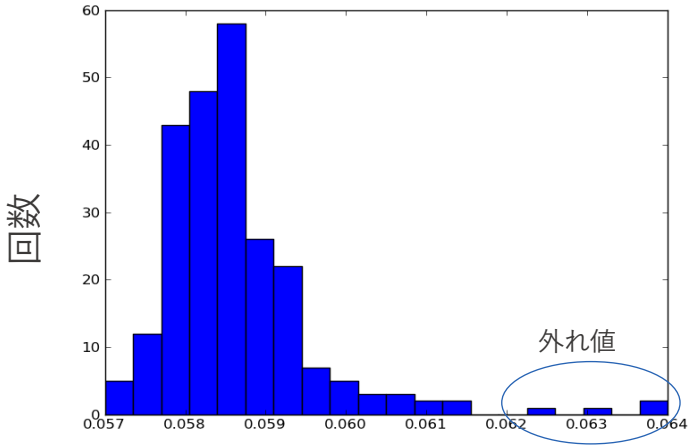
例: ７月の１時間あたりの電力使用料

- ふたこぶ型の分布になっている
- 原因は夜と昼の電力使用量の性質が異なるためである。
- 昼と夜で層別が必要

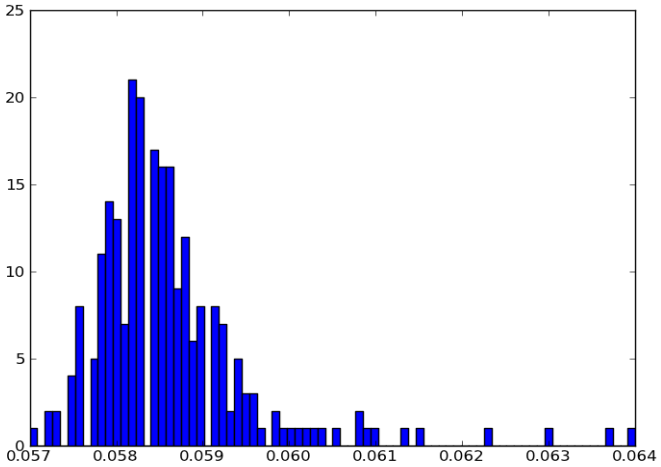


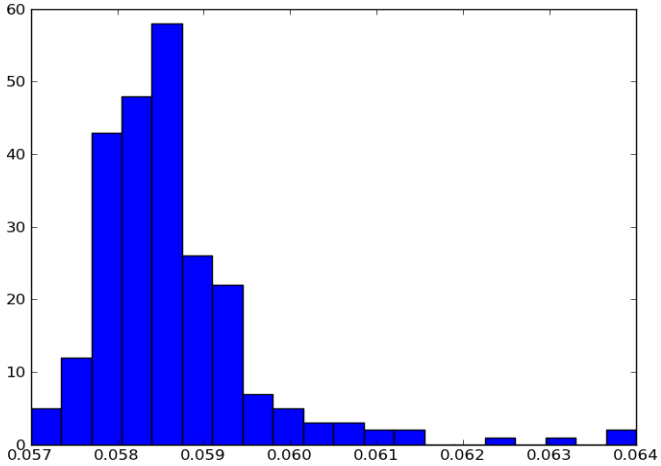
例: 7月の新宿で観測された放射線量

- 基本的には正規分布ではあるが、外れ値が幾つか見られる。
- 外れ値がなぜ起こったか究明することが必要。



作り方の悪いヒストグラム





グラフが歯抜けしているので良くない.
区間の設定(階級の幅)が不適切.

標準偏差と分布の関係

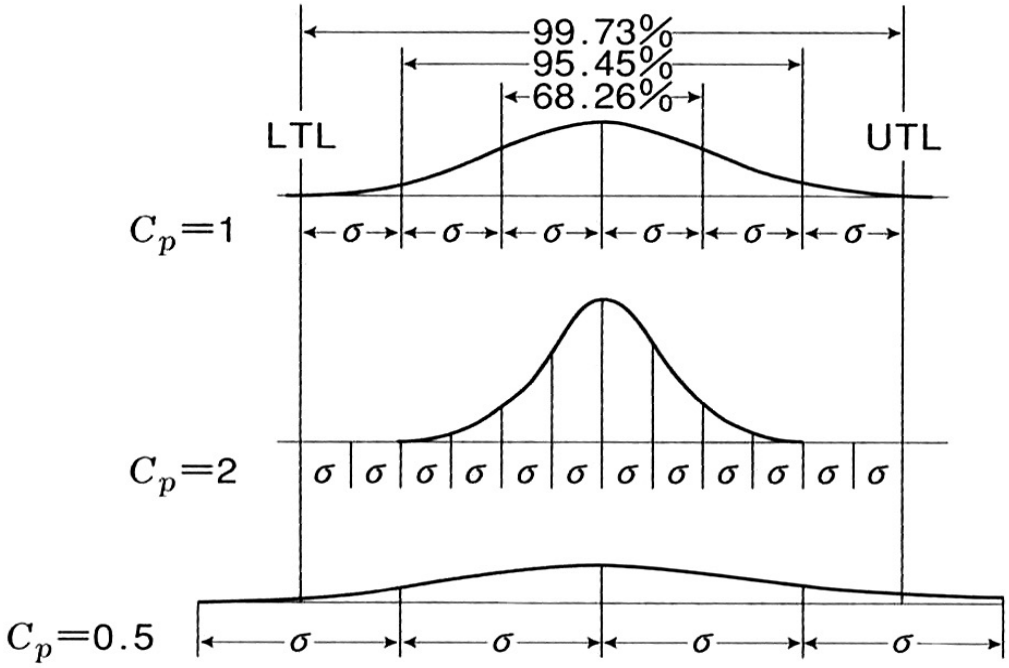
- 分散

$$V = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x})^2$$

- 標準偏差

$$S = \sqrt{V}$$

$$(\sigma)$$



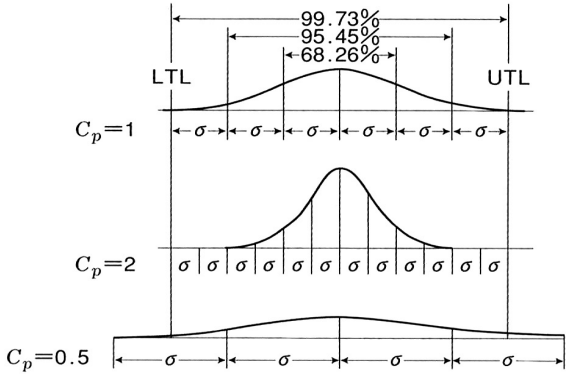
（QC数学の話 大村平より）

工程能力指数

- 製品規格と分布の関係を表す指標
- 大きければ大きいほどよいが、大きすぎる場合過剰に対策をしている場合もある

- 上方許容限界(UTL)
  - 品質の上方限界
- 下方許容限界(LTL)
  - 品質の下方限界
- 工程能力指数

$$C_p = \frac{\mathrm{UTL} - \mathrm{LTL}}{6\sigma}$$



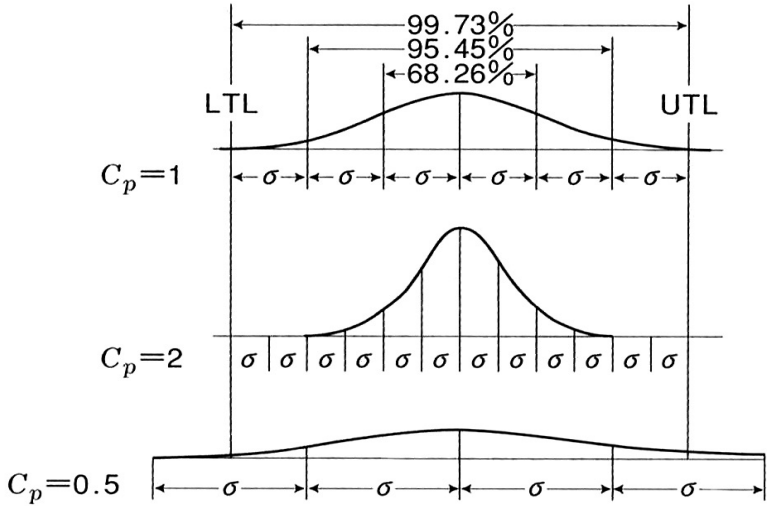
（QC数学の話 大村平より）

工程能力指数は製品の分布が規格内にどれくらい収まっているかという指標

Cp=1なら生産された全製品のうち99.73%が規格内に収まっている。

学力偏差値

- (得点ー平均点)÷標準偏差x10+50
- で計算される.
- これは，テストの結果を平均点を50点にし，標準偏差を10点に正規化したという意味になる.



おまけではあるが，統計についてはこれさえ分かれば良い！！

- 統計量には主観が入る．
  - 統計量は，統計量を計測する側が想定したモデルのもとに計算される．
    - 例：平均や分散をとるということは，それを計算する時点でデータが正規分布をしていると想定している．つまり，データが正規分布していなければ意味がないものである．
  - モデルは主観的に想定するため，統計量には主観が入る．

- データを目視せよ．
  - より正確にモデルを選ぶ（主観の精度をあげる）ためにデータの分布の形状を確認する．

- All models are wrong, but some are useful「全てのモデルは間違っている，しかし中には使えるものもある．」(Box)
  - よく使われる統計量や統計手法を絶対だと思わないこと！！
  - データを多面的に見て（様々なモデルを想定して）判断すべきである．

- 人工知能的なものは統計手法が使われているので，上記のことは人工知能的なものにも当てはまる．

演習

演習

- システム開発の進捗管理やソフトウェアの品質管理などで用いられるPDCAサイクルの“P”，“D”，“C”，“A”は，それぞれ英単語の頭文字をとったものである。3番目の文字“C”が表す単語はどれか。（基本情報技術者平成26年春期）

1. Challenge
2. Change
3. Check
4. Control

演習

- システム開発の進捗管理やソフトウェアの品質管理などで用いられるPDCAサイクルの"P"，"D"，"C"，"A"は，それぞれ英単語の頭文字をとったものである。3番目の文字"C"が表す単語はどれか。基本情報技術者平成26年春期

1. Challenge
2. Change
3. **Check**
4. Control

演習

- 情報セキュリティマネジメントがPDCAサイクルに基づくとき，Cに相当するものはどれか。(ITパスポート平成30年春期)

1. 情報セキュリティの目的，プロセス，手順の確立を行う。
2. 評価に基づいた是正及び予防措置によって改善を行う。
3. プロセス及び手順の導入，運用を行う。
4. プロセスの効果を測定し，結果の評価を行う。

演習

- 情報セキュリティマネジメントがPDCAサイクルに基づくとき，Cに相当するものはどれか。(ITパスポート平成30年春期)

1. 情報セキュリティの目的，プロセス，手順の確立を行う。
   Planに当てはまります.
2. 評価に基づいた是正及び予防措置によって改善を行う。
   Doに当てはまります.
3. プロセス及び手順の導入，運用を行う。
   Actに当てはまります.
4. **プロセスの効果を測定し，結果の評価を行う。**

演習

- ヒストグラムを説明したものはどれか．（基本情報平成22年秋期）

1. 原因と結果の関連を魚の骨のような形態に整理して体系的にまとめ，結果に対してどのような原因が関連しているかを明確にする。
2. 時系列的に発生するデータのばらつきを折れ線グラフで表し，管理限界線を利用して客観的に管理する。
3. 収集したデータを幾つかの区間に分類し，各区間に属するデータの個数を棒グラフとして描き，ばらつきをとらえる。
4. データを幾つかの項目に分類し，出現頻度の大きさの順に棒グラフとして並べ，累積和を折れ線グラフで描き，問題点を絞り込む。

演習

- ヒストグラムを説明したものはどれか．（基本情報平成22年秋期）

1. 原因と結果の関連を魚の骨のような形態に整理して体系的にまとめ，結果に対してどのような原因が関連しているかを明確にする。
   特性要因図です．
2. 時系列的に発生するデータのばらつきを折れ線グラフで表し，管理限界線を利用して客観的に管理する。
   管理図です．
3. **収集したデータを幾つかの区間に分類し，各区間に属するデータの個数を棒グラフとして描き，ばらつきをとらえる。**
4. データを幾つかの項目に分類し，出現頻度の大きさの順に棒グラフとして並べ，累積和を折れ線グラフで描き，問題点を絞り込む。
   パレート図です．