SQC勉強会(3)
ヒストグラムと代表値

本の宣伝

- この勉強会の元本です
- サンプルデータもこの本からとっています
- 著者の先生もいい人です
- わかりやすくコンパクトな本でお勧めです

(講演を聞いただけですが、人のよさそうなひとでした)

- フリーソフトウェアRによる統計的品質管理入門
- 編著：荒木孝治
- 出版社：(株) 日科技連出版社

本日の内容

- ヒストグラム
- 代表値と要約統計量
- 分布(ヒストグラムからモデルへ)

品質の管理とは？

- 品質:「(モノの)利用における(効用の)適合度」
(Montgomery, 2005)
- 4要素
  - 対象
    - 何(部門、工程)を対象とするのか
  - 特性
    - 対象のどのような性質を管理したいのか
  - 標準(水準)
    - 許容できる特性の測定量はどの範囲か？
  - 要因
    - 特性の計量値に影響を与える要素は何か？
- SQCでは特に標準の決め方と活動データと標準との関係の判断に関する手法を重視

QC7つ道具とは

2010年1月14日

- 定量データを基にしてSQCを行う際の基本的ツール
- 視覚的に
  - 課題の整理
  - 問題要因の特定・標準の設定
  - 対策実施後の品質の判断

※新QC7つ道具：主に定性的なものに対して適用

QC7つ道具とは？

2010年1月14日

- グラフ・管理図：データの傾向を知る 勉強会で扱いません
- チェックシート：手順のチェック
- パレート図：重点対策する要因を決める
- ヒストグラム：データの特性を知る
- 特性要因図：バラツキの要因の解析
- 散布図：要因間の関係を見る
- 層別：要因の存在を調べる

テーマ：「代表値」をどのように決めるのか

2010年1月14日

- 代表値：「グループ」(群)を代表する値
- グループの全体的な様子を1つの値に「代表」させる
- よく使うのは「平均値」
- しかし、平均値を用いるとグループを代表しない場合も・・・
- いくつかの代表値を学習し、適切なものを選択する
- 鉄則：事前にヒストグラムを描いて判断する

ヒストグラム

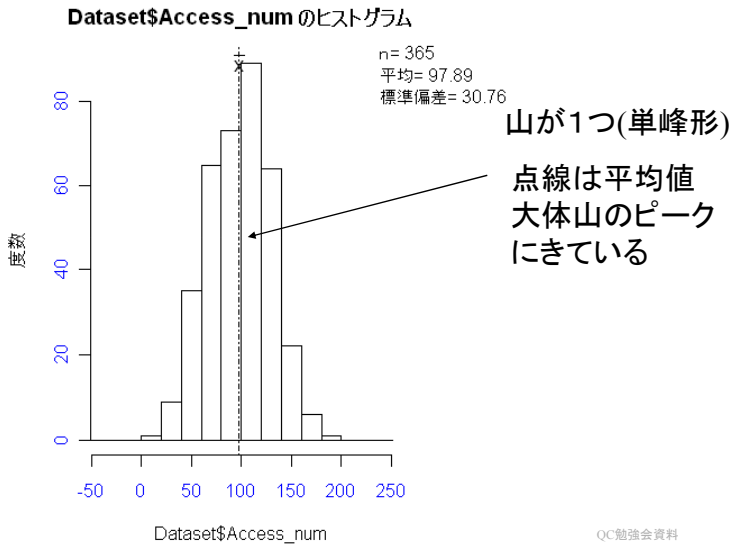
2010年1月14日

- ヒストグラム
  - タテに値(区間)、ヨコに発生回数(頻度)をとった棒グラフ
- あるシステムの1日あたりのアクセス回数のデータ(架空)
- Access_num.csv

ヒストグラム

- データの読み込み
- 「データ」→「データのインポート」
- Macの人はクリップボードからコピー＆ペーストしたほうがよい（らしいです）
- 「QCツール」→「QCヒストグラム」

ヒストグラム

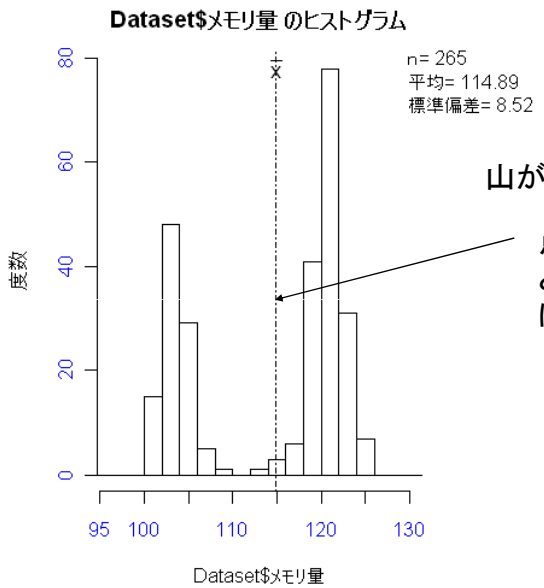


山が1つ(単峰形)

点線は平均値
大体山のピーク
にきている

本当に平均値は代表値？

- 現実の場合は必ずしもそうでない場合も
- あるシステム(サーバ2台構成)のメモリ使用量の推移
- 全体のヒストグラムを見てみる

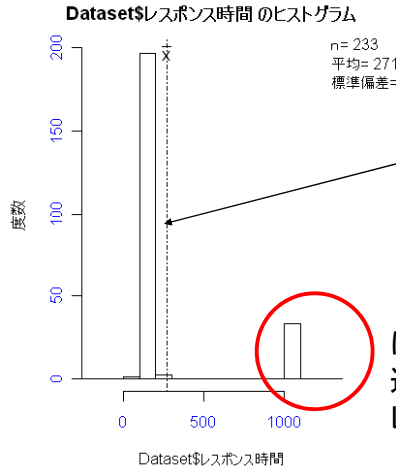


山が2つ(複峰形)

点線は平均値
どちらの山のピーク
にも来ていない

本当に平均値は代表値？

- あるシステムのレスポンス時間(ミリ秒)
- 2台構成のうち、途中で1台がハード故障
- その間は1台のみで運用
- response.csv

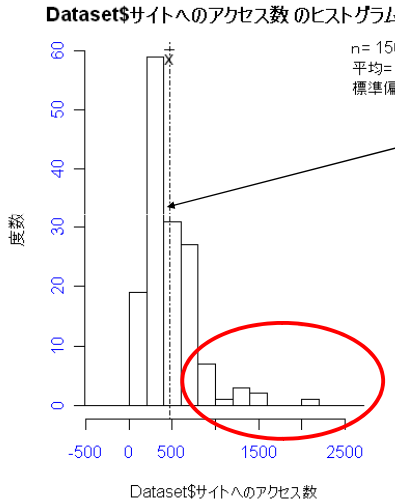


点線は平均値
山のピークでは
ない

はずれ値：1台のみでの
運用をしたことによる
レスポンスの低下

本当に平均値は代表値？

- あるECサイトの1日あたりの来訪者
- Access_num2.csv



点線は平均値
山のピークから
微妙にずれる

右にすそが長い
いくつかの日はアクセス
数が非常に多かった

本当に平均値は代表値？

- ヒストグラムを見て、
  - 単峰：山が1つ
  - 対称：左右が対象

の場合は平均値と山のピークは一致

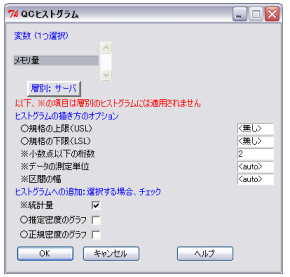
平均値＝代表値

- それ以外のときは別の手立てを考える

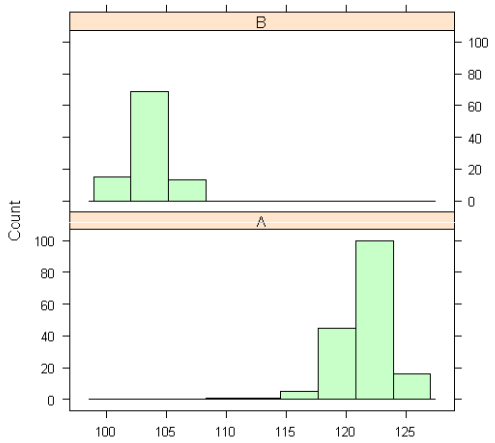
QC(データ解析)でよく用いられる手法

その1：層別

- memory.csvを再度見てみる
- サーバ情報がデータに存在
- サーバによってメモリ量に変化はないか？
- 層別の実施



層別のヒストグラム

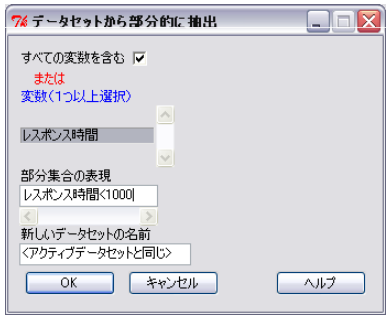


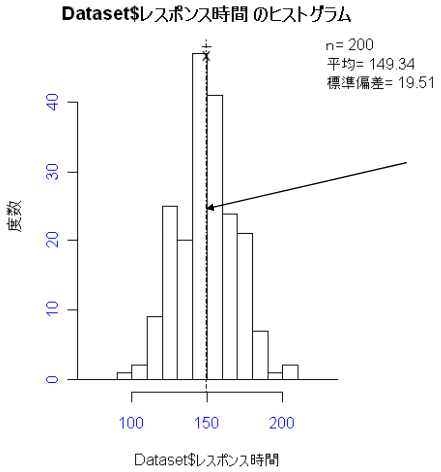
- 層別の結果
- 2つの山の原因は
サーバの種別によるもの
- サーバごとに平均値を出し、代表値とみなすのが合理的

QC(データ解析)でよく用いられる手法

その2：はずれ値を取り除く

- response.csvを見てみる
- 明らかに1000ミリ秒を超えたものは少し異常？
- システム設計上そうだった場合はこれをはずれ値とみなして取り除く





点線は平均値
山のピークとほぼ
一致

QC(データ解析)でよく用いられる手法

その3：他の代表値を使う

→困ったらまず中央値を見てみる

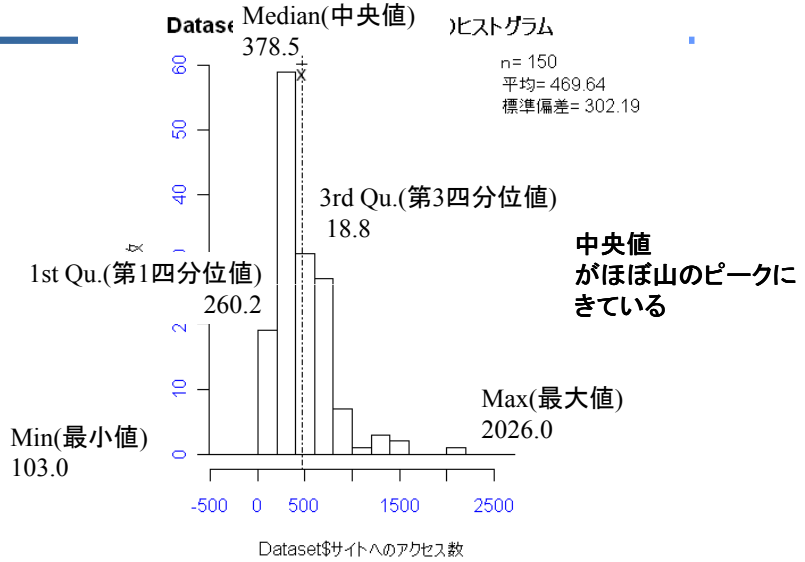
ヒストグラムの形に影響を受けない（頑健性）

- 平均値
  - (すべての観測値の和)／(観測数)
- 中央値(メディアン)
  - すべての観測値を昇順(降順)に並べたときの真ん中の値
- 最頻値（モード）
  - すべての観測値でもっとも発生回数が高い値

統計的に扱いやすい

中央値の適用

- Access_num2.csvを見てみる
- 「統計量」→「要約」→「アクティブデータセット」で色々な代表値が出る



中央値
がほぼ山のピークに
きている

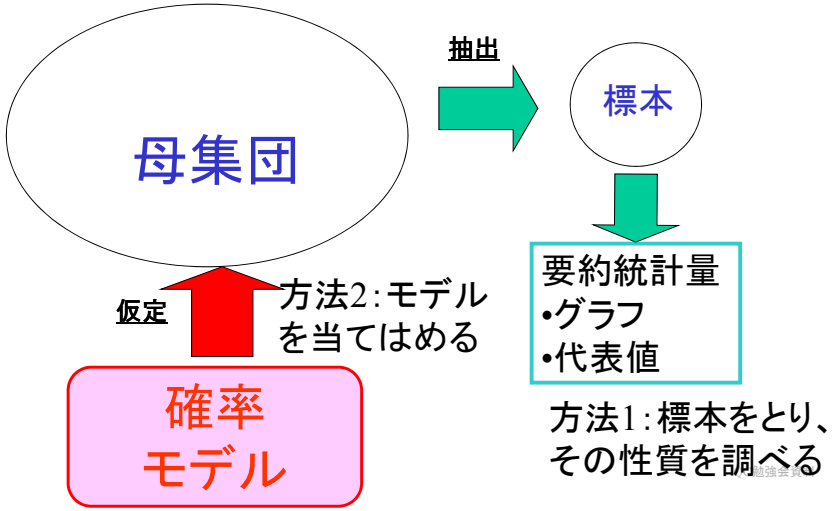
山のピーク＝最頻値ではないの？

- 概略的にはそのとおりです。
- しかし、通常は
  - 観測データ(連続量)の場合、全く同じ値が2度でることはほとんどない
  - データをある位で丸めると恣意性が入ってしまう
  - 後で出てくる検定(統計分析)が適用しづらい

  ため、中央値までの適用にとどめます。

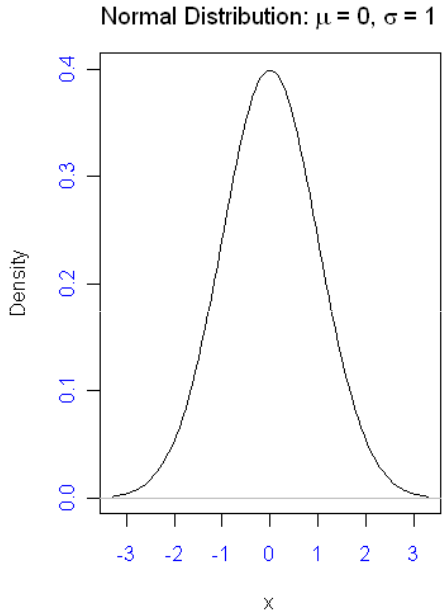
ヒストグラムからモデルへ

母集団の性質を理解するためにどうするのか？



確率モデルとは？

- 確率
  - ある事象(値)が発生する割合
- 確率密度関数
  - ヨコに事象(値)、タテに確率(割合)を採ったグラフ
- 「分布」→「正規分布」→「正規分布を描く」



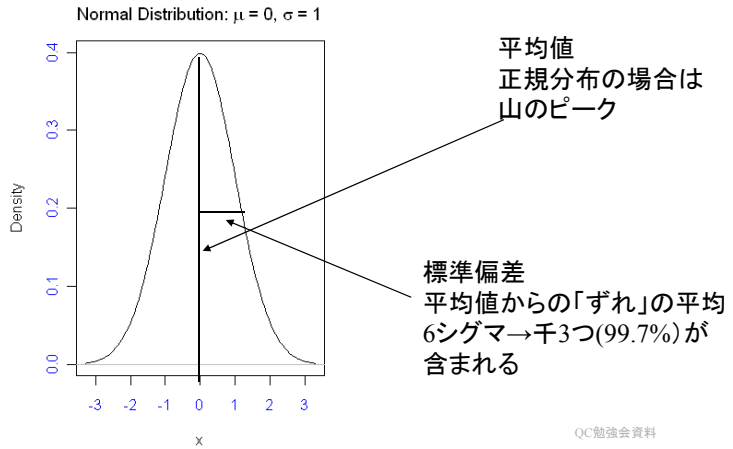
ヒストグラムと確率密度関数

- ヒストグラム
  - ある事象(値)に対する実際の発生回数(頻度)
- ヒストグラムの発生回数を多くし、割合に直すと確率密度関数になる

(頻度論的な考え方)

正規分布:重要な分布

- 正規分布:N(平均値, 標準偏差)



次回予告

- 正規分布を覚えて活用しよう
  - シックスシグマ(の概論)
  - 管理図