データから情報を集約し戦略的に活用するための
データ分析/統計解析/マイニングソフトウエア

S-PLUS

株式会社 数理システム

S-PLUS**グループ**

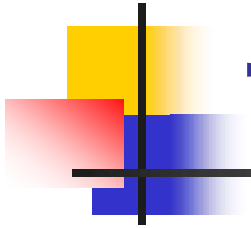
田澤 司(tazawa@msi.co.jp)

電子的データの時代

- 毎日データを記録されて生きる時代 -

- 商取引データ（posデータ、銀行ATMデータ、カード会員履歴データ、suica&pasmo等、通話履歴データ…）
- WEB等のアクセスデータ（ログ）
- 市況データ（証券市場、為替市場…）
- 生産履歴データ（工場データベース）
- 年金データとか？

何かの行動に伴い、毎日データが発生し、データベースに記録されていく



データベースは大規模になりました

- 大量の過去データからの容易な検索
  - 1件1件のトレイサビリティは大幅に向上
- 事象の全体的傾向/変化を把握できないか？
  - 「個」から「全体」へ
- DB単体では「データ集計」まで
  - 多くの要因が複合条件下で事象を生起する
  - 大規模データにはそれなりの分析手法が必要

「集計」から「分析」へ

(本日のテーマ)

データから有益な情報を得よう

- 大量の業務実績データは情報の宝庫
  - 取引実績データ,顧客属性データ,製造実績データ…
- 分析により「現状の正しい認識」が可能
- 分析により「将来の予測」が可能
  - 将来に備えたアクションが取れる

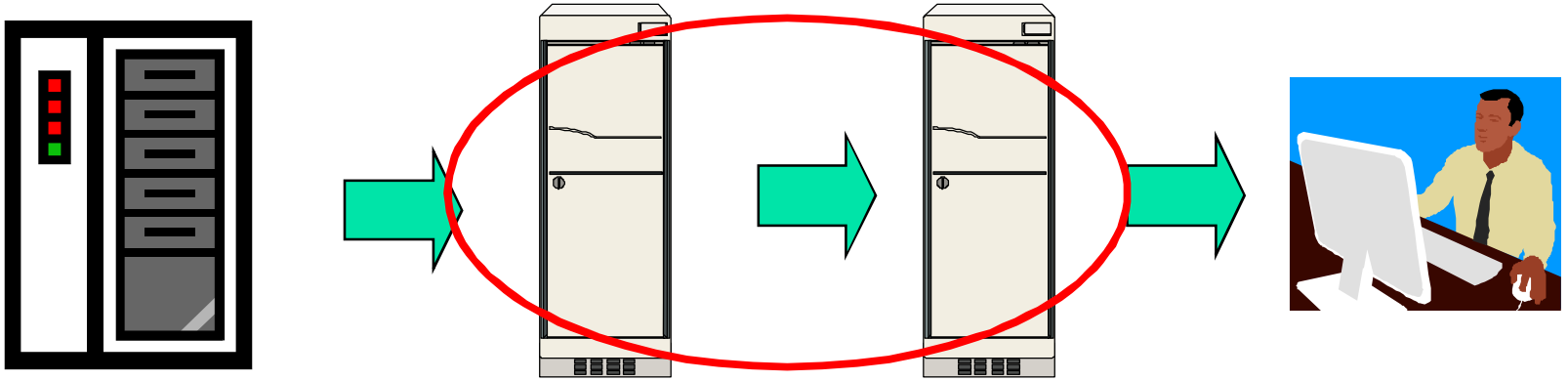
従来にない結果を出してこそ大規模DBの意義

DBサーバ
(原データ)

分析サーバ
(データ加工)

WEBサーバ
(配信)

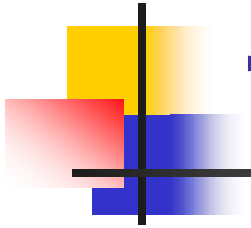
ユーザ端末
(表示)



この辺りのお話

具体的用途の一例

- 売上に寄与する原因は何か解明
  - 顧客の属性や周辺環境と実際の売上との関連性分析
  - WEBログなどの分析
- 需要や市場動向の高度な分析、将来予測
  - 各事業単位での将来の需要予測、売上予測
  - 時系列解析により将来を予測、政策決定に役立てる
- 金融業での最先端技術を利用した資産運用（金融工学）
  - 大量の選択肢（株や債権）から、ローリスクハイリターンの組合せを取捨選択
- 工場での収集データからの品質管理、工程管理
  - ルーチンワークの自動化による省力化、分析自体の高度化



データの分析

- データの可視化により、大規模データを短時間に把握可能
- データマイニング/統計解析手法により、複雑に相関したデータの奥に潜む法則を解明

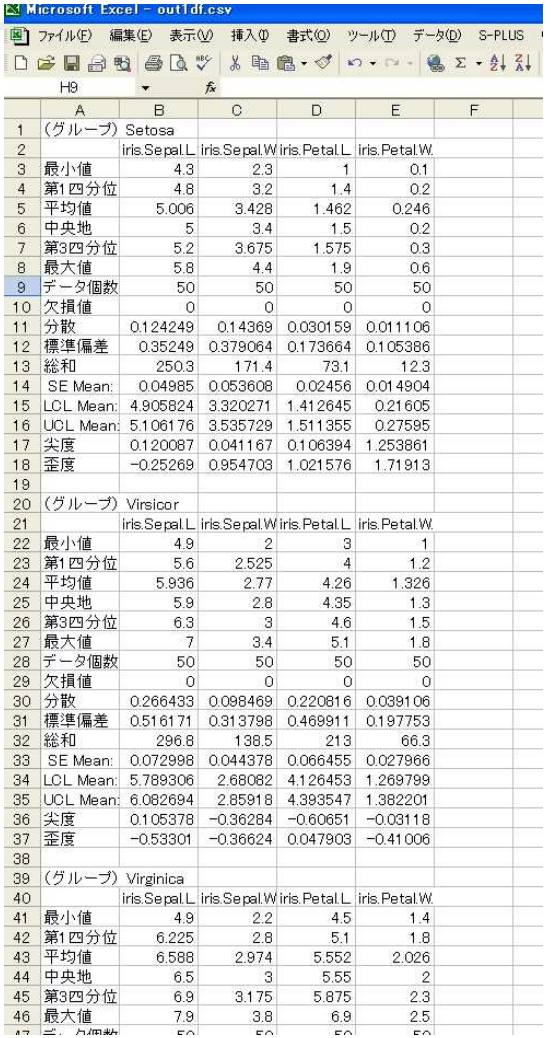
今回紹介する分析系製品

- 統計解析/グラフソフト S-PLUS
  - 実績ある本格派解析ソフト&開発環境
- ネットワーク対応分析サーバ
  S-PLUS Enterprise Server
  - ネットワークの中で高度分析サービスを提供

データの可視化は便利

（従来の主流）
分析結果の大量の表出力

- 精査する気になりますか？
- 理解、納得できますか？
- 短時間で把握できますか？

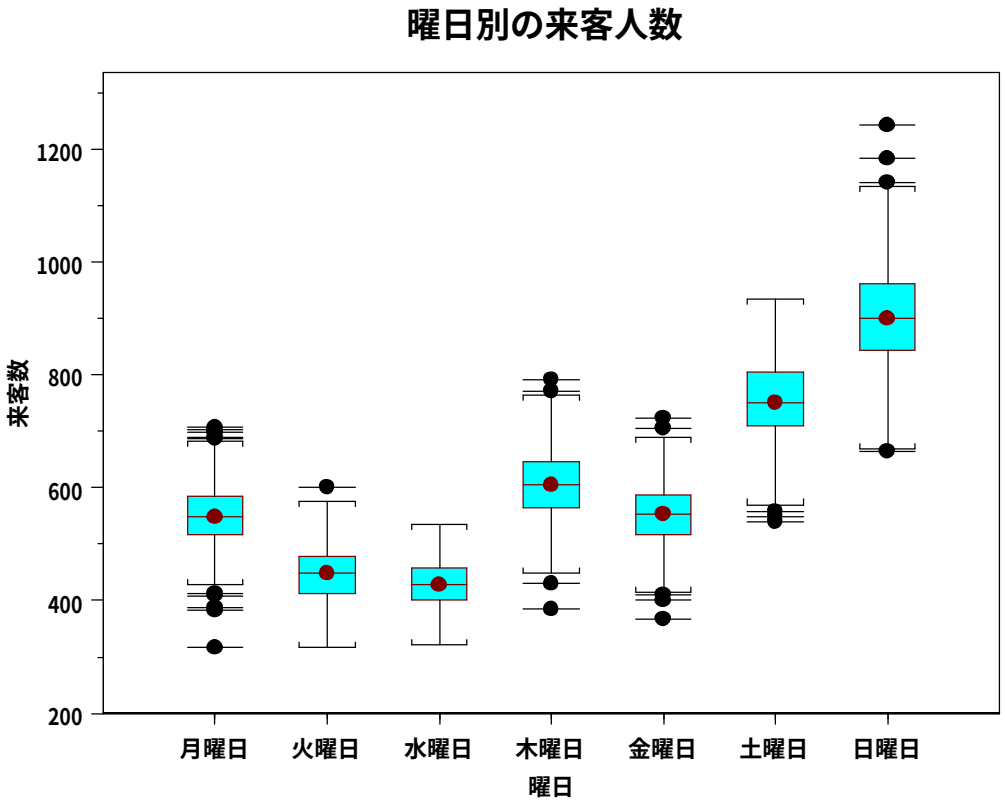

Microsoft Excel - outIdf.csv

ファイル(F) 編集(E) 表示(V) 挿入(I) 書式(O) ツール(T) データ(D) S-PLUS

H9

	A	B	C	D	E	F
1	(グループ)	Setosa				
2		iris.Sepal.L	iris.Sepal.W	iris.Petal.L	iris.Petal.W	
3	最小値	4.3	2.3	1	0.1	
4	第1四分位	4.8	3.2	1.4	0.2	
5	平均値	5.006	3.428	1.462	0.246	
6	中央地	5	3.4	1.5	0.2	
7	第3四分位	5.2	3.675	1.575	0.3	
8	最大値	5.8	4.4	1.9	0.6	
9	データ個数	50	50	50	50	
10	欠損値	0	0	0	0	
11	分散	0.124249	0.14369	0.030159	0.011106	
12	標準偏差	0.35249	0.379064	0.173664	0.105386	
13	総和	250.3	171.4	73.1	12.3	
14	SE Mean:	0.04985	0.053608	0.02456	0.014904	
15	LCL Mean:	4.905824	3.320271	1.412645	0.21605	
16	UCL Mean:	5.106176	3.535729	1.511355	0.27595	
17	尖度	0.120087	0.041167	0.106394	1.253861	
18	歪度	−0.25269	0.954703	1.021576	1.71913	
19						
20	(グループ)	Virsicor				
21		iris.Sepal.L	iris.Sepal.W	iris.Petal.L	iris.Petal.W	
22	最小値	4.9	2	3	1	
23	第1四分位	5.6	2.525	4	1.2	
24	平均値	5.936	2.77	4.26	1.326	
25	中央地	5.9	2.8	4.35	1.3	
26	第3四分位	6.3	3	4.6	1.5	
27	最大値	7	3.4	5.1	1.8	
28	データ個数	50	50	50	50	
29	欠損値	0	0	0	0	
30	分散	0.266433	0.098469	0.220816	0.039106	
31	標準偏差	0.516171	0.313798	0.469911	0.197753	
32	総和	296.8	138.5	213	66.3	
33	SE Mean:	0.072998	0.044378	0.066455	0.027966	
34	LCL Mean:	5.789306	2.68082	4.126453	1.269799	
35	UCL Mean:	6.082694	2.85918	4.393547	1.382201	
36	尖度	0.105378	−0.36284	−0.60651	−0.03118	
37	歪度	−0.53301	−0.36624	0.047903	−0.41006	
38						
39	(グループ)	Virginica				
40		iris.Sepal.L	iris.Sepal.W	iris.Petal.L	iris.Petal.W	
41	最小値	4.9	2.2	4.5	1.4	
42	第1四分位	6.225	2.8	5.1	1.8	
43	平均値	6.588	2.974	5.552	2.026	
44	中央地	6.5	3	5.55	2	
45	第3四分位	6.9	3.175	5.875	2.3	
46	最大値	7.9	3.8	6.9	2.5	

S-PLUSによるデータの可視化の例(1)
(あるスーパーの曜日別来店人数)



曜日別の傾向、バラツキを一目で把握
->予測へ

S-PLUSによるデータの可視化の例(2)

(アンケート調査によるマーケティングリサーチ)

- ATT社とOCC社はライバル関係にある電話会社
- アンケート調査により、顧客の選択理由を探る

Microsoft Excel - market.survey.csv

ファイル(F) 編集(E) 表示(V) 挿入(I) 書式(O) ツール(T) データ(D) S-PLUS ウィンドウ(W) ヘルプ(H) Adobe PDF(B)

MS Pゴシック

O10

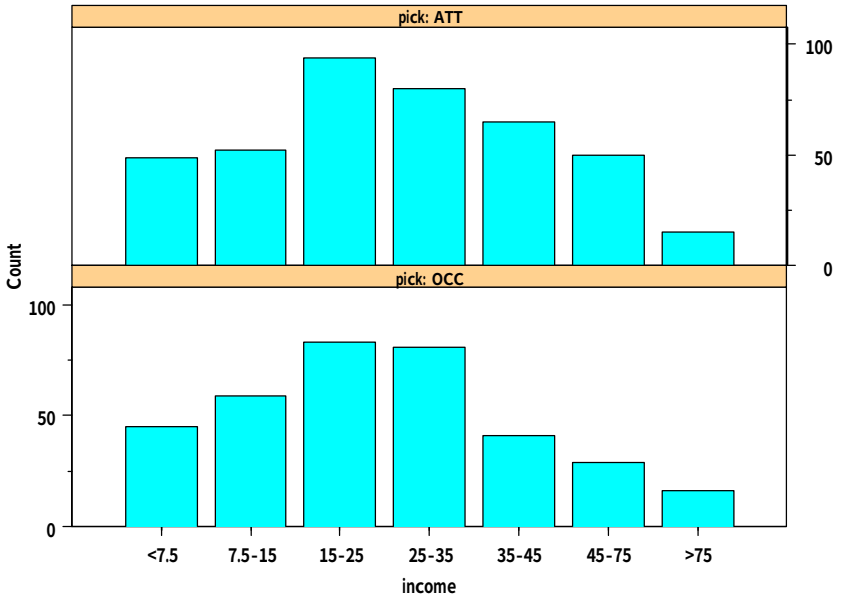
	A	B	C	D	E	F	G	H	I	J	K
1	顧客ID	選択会社	年収	引越頻度	年齢	学歴	職業	電話利用頻度	電話帳	サービス	カード
2	1001	OCC	<7.5	0	35-44	HS	F	9	Y	N	N
3	1002	ATT	<7.5	2	25-34	HS	F	5	Y	N	N
4	1003	ATT	<7.5	0	35-44	<HS	H	1	Y	N	N
5	1004	ATT	<7.5	0	55-64	HS	F	9	N	N	N
6	1005	OCC	<7.5	0	65+	<HS	R	5	N	N	N
7	1006	OCC	<7.5	0	65+	HS	D	1	N	N	Y
8	1007	OCC	<7.5	0	55-64	<HS	R	1	N	N	Y
9	1008	ATT	<7.5	1	25-34	HS	F	3	N	N	N
10	1009	OCC	<7.5	0	65+	<HS	R	5	N	N	N
11	1010	ATT	<7.5	2	18-24	HS	U	26	Y	N	N
12	1011	OCC	<7.5	0	45-54	HS	F	60	N	N	N
13	1012	OCC	<7.5	0	65+	<HS	R	55	N	N	N
14	1013	OCC	<7.5	0	65+	<HS	R	0	N	N	N
15	1014	OCC	<7.5	0	55-64	<HS	H	0	Y	N	N

(元データ)

全く同じ分析ですが…

統計学を知らない人に、どちらが説明しやすい？

ATT社v.s.OCC社の選択において年収によって傾向の違いがあるか？



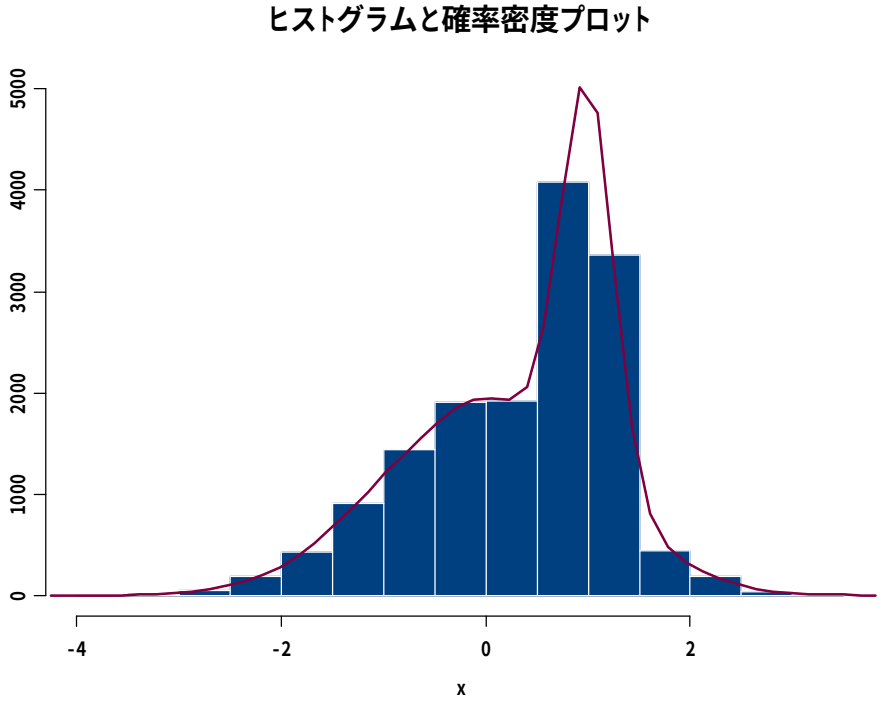
pick	income							
	<7.5	7.5-15	15-25	25-35	35-45	45-75	>75	RowTotl
OCC	47	62	90	90	42	29	16	376
	0.12	0.16	0.24	0.24	0.11	0.077	0.043	0.48
	0.49	0.54	0.49	0.53	0.39	0.36	0.5	
	0.06	0.079	0.11	0.11	0.054	0.037	0.02	
ATT	49	52	95	81	65	51	16	409
	0.12	0.13	0.23	0.2	0.16	0.12	0.039	0.52
	0.51	0.46	0.51	0.47	0.61	0.64	0.5	
	0.062	0.066	0.12	0.1	0.083	0.065	0.02	
ColTotl	96	114	185	171	107	80	32	785
	0.12	0.15	0.24	0.22	0.14	0.1	0.041	

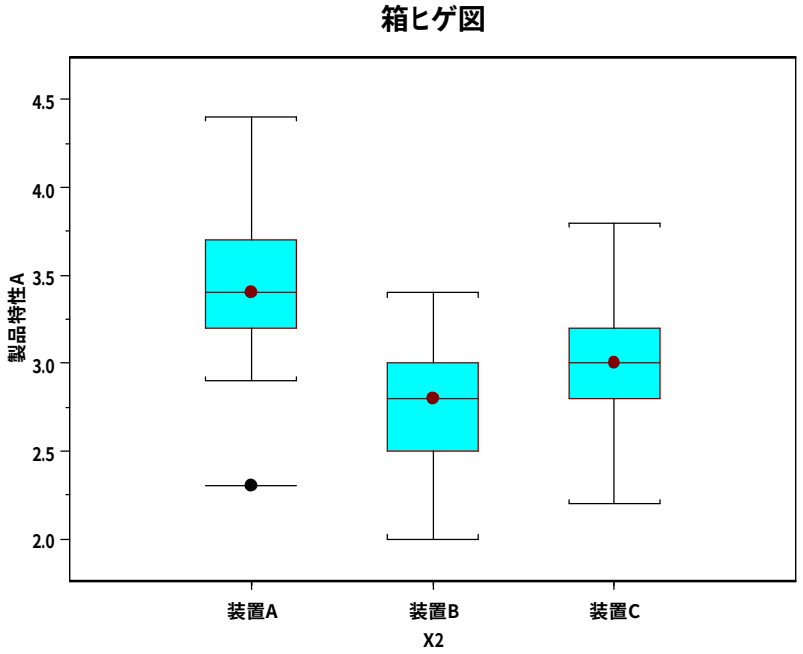
Test for independence of all factors
Chi^2 = 11.15405 d.f.= 6 (p=0.08373029)
Yates' correction not used

ビジュアリゼーションの説得力！

S-PLUSによるデータの可視化の例(3)
(工場の製造データで品質管理)

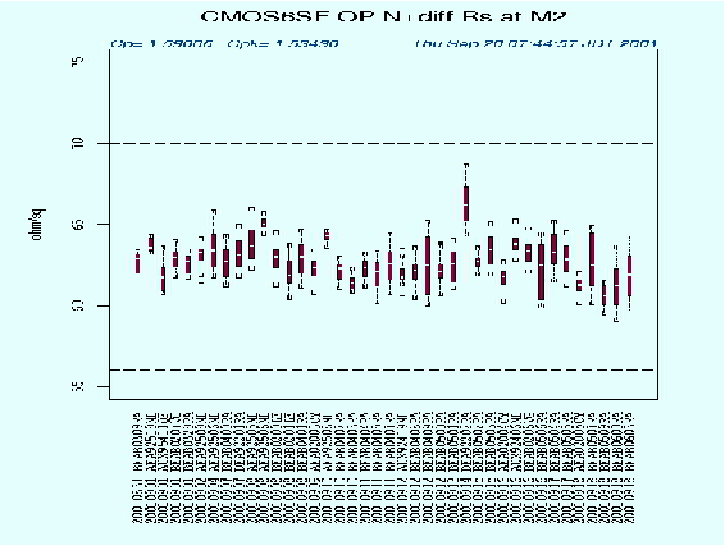
- 中央値、バラツキなど1変数の概要を見る
  - ヒストグラムやボックスプロット (箱ヒゲ図)



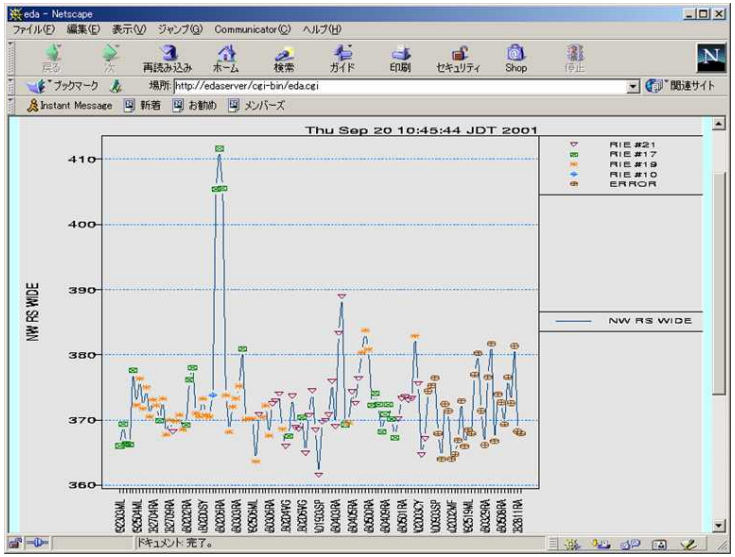


S-PLUSによる実稼動システムの例 (半導体工場でのデータの可視化)

- **時系列的推移 (トレンド) を見る**



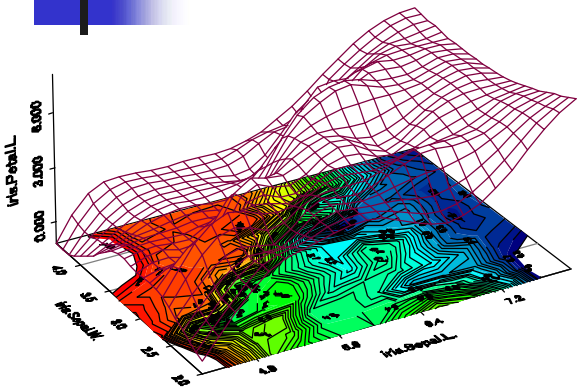
製品電気特性分布のロット毎時系列推移

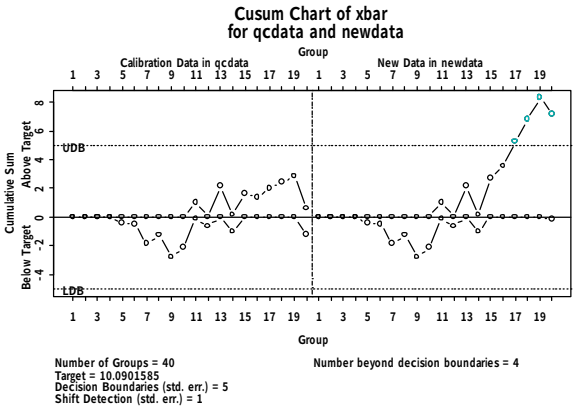


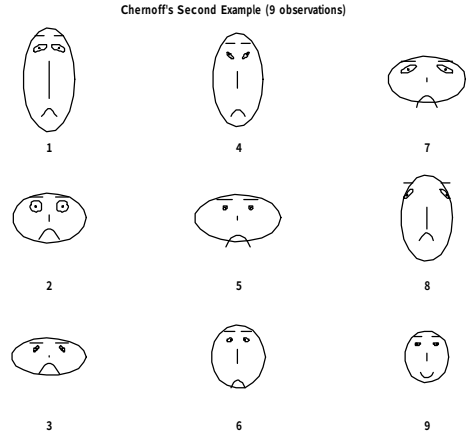
イントラネットで条件設定/データ抽出を行い描画したグラフ

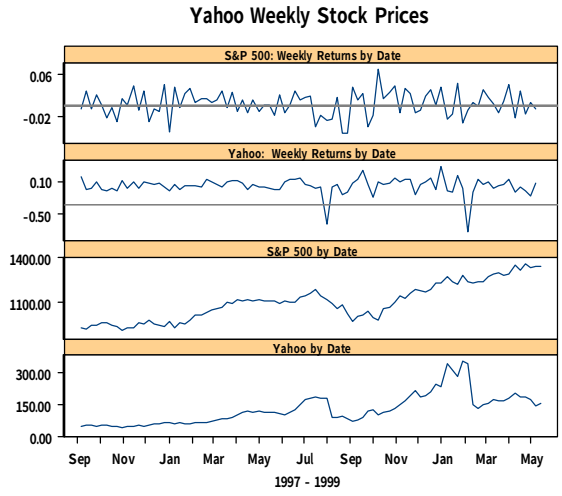
WEBにより誰でも簡単に閲覧

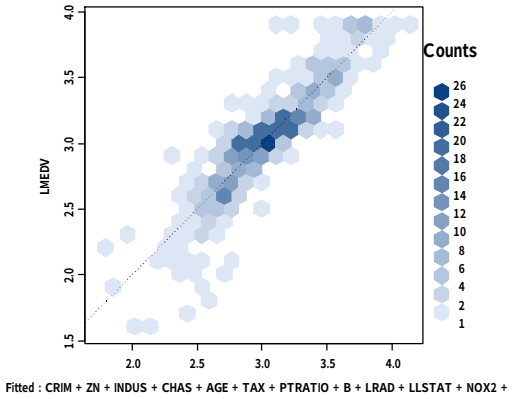
数々のグラフ手法を搭載
(クリックで容易に利用可能)

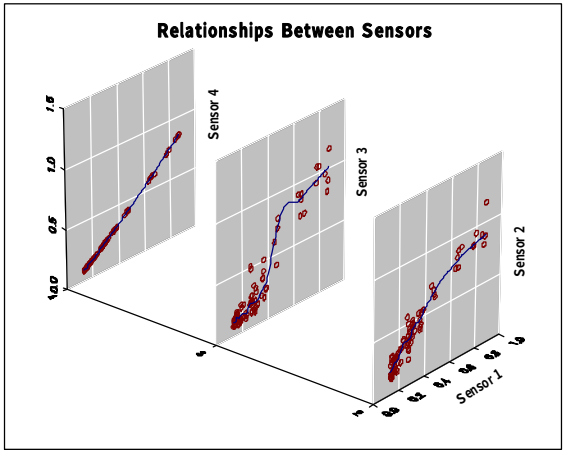


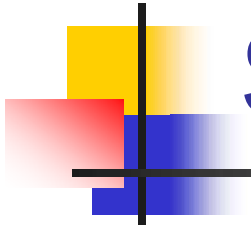










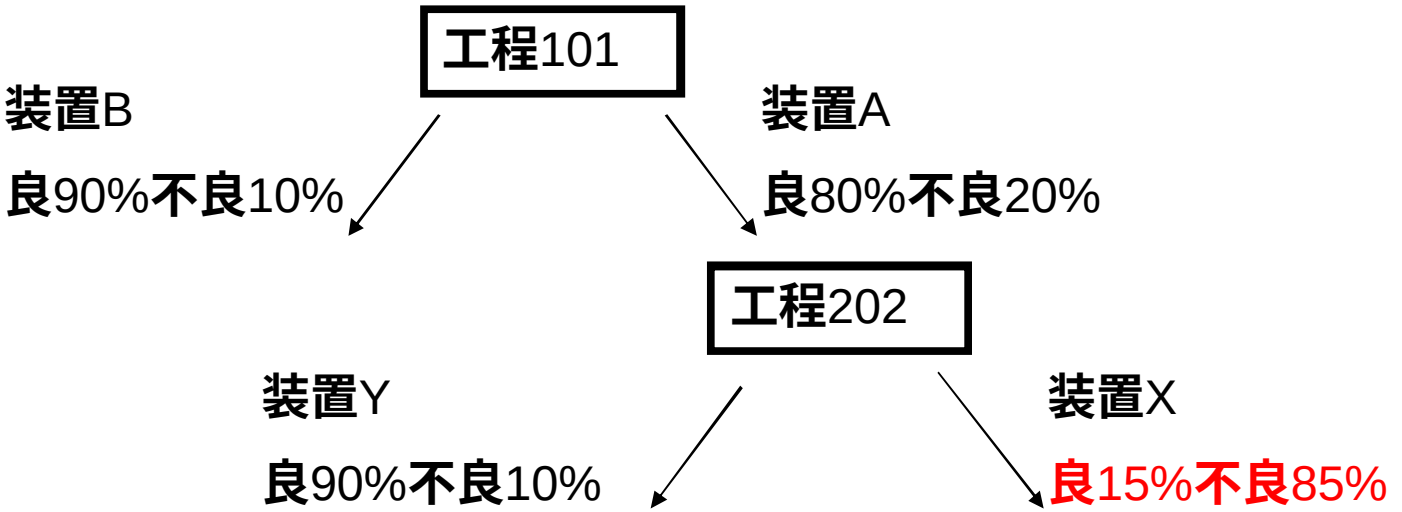


S-PLUSの機能

- 柔軟なデータ加工機能
  - 欠損値処理やデータの補正が楽々
- 大規模データ対応 (Enterprise版)
- 豊富な統計解析/マイニング機能
  - ほとんど全ての分野を網羅
- 本格的なプログラミング機能
  - 自動化による合理化
- DB等とのインターフェース

解析手法の一例

- **決定木** (Decision Tree)

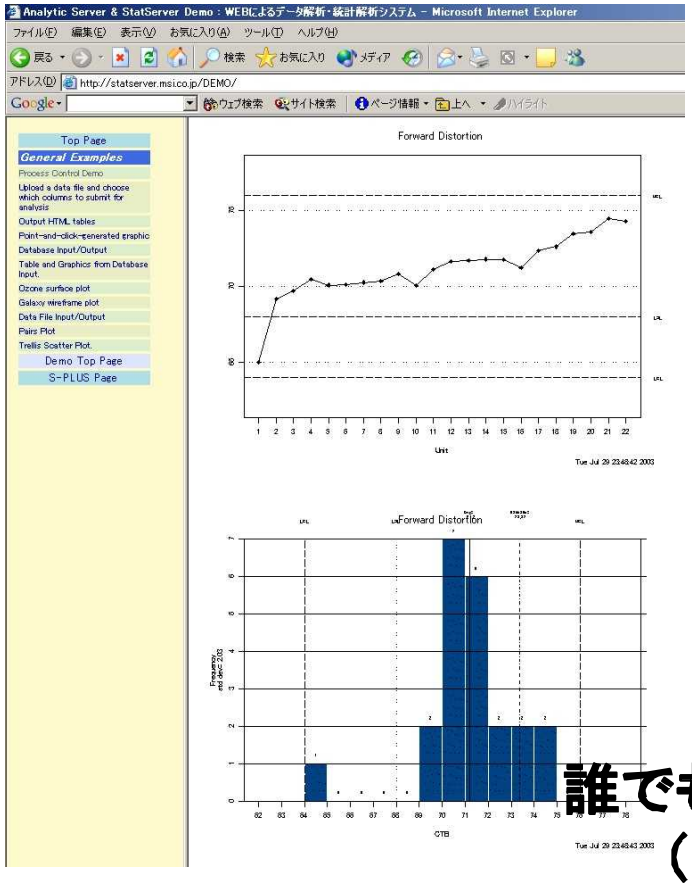


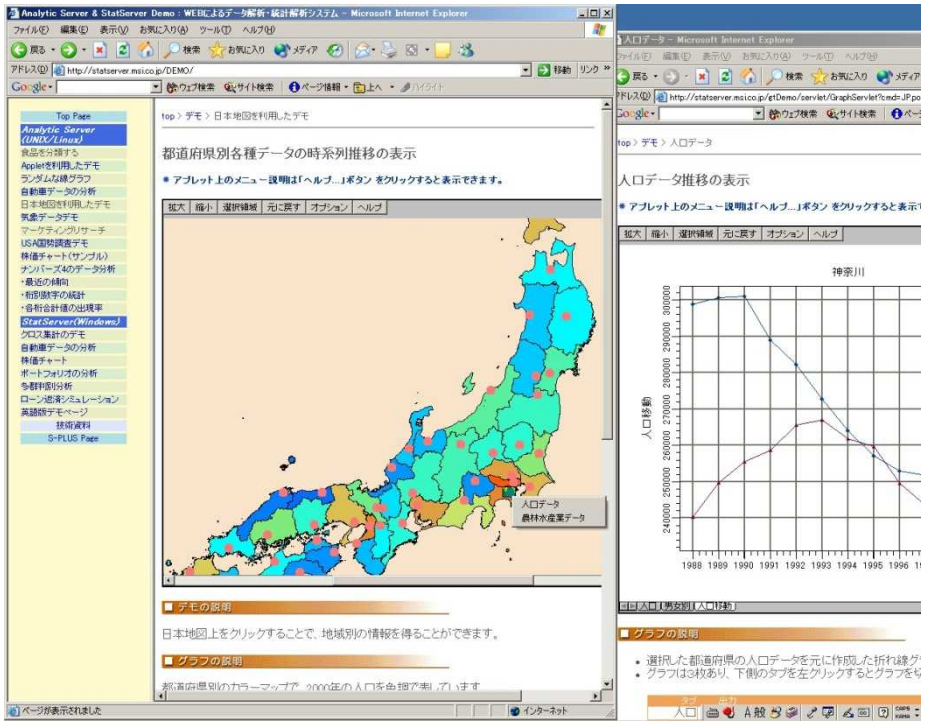
工程101で装置Aで処理され、なおかつ工程202で装置Xで処理された製品のみ著しく歩留りが悪い (複合条件での問題の検出)

S-PLUS Enterprise Server

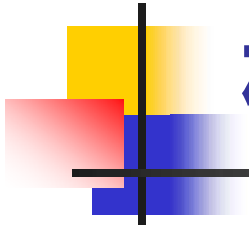
- ネットワークでの高度解析サービス
- 解析用WEBサーバとしての構築も可能
- 利用者はブラウザやExcel、VB/Javaプログラムなどから簡単に利用
  - 使い慣れたGUIから高度な解析
- 定型サービスを広汎に提供可能
  - 例）ルーチンな品質管理、工程管理図
  - 例）ルーチンな各支社の販売レポートなど

S-PLUS Enterprise Server適用例





誰でも簡単に大規模データの分析が出来ます
（高機能なポータルサイトの運営にも）



株式会社 数理システムについて

- 独立系ソフトウエア開発/コンサル
  - 1982年設立、社員数 70名 (うち8割はエンジニア)
- ソフトウエア開発に軸足
  - 特にUNIX/Linuxベースでは20年以上の実績
  - インターネット技術も長い実績
- 数理科学系ソフトウエアを専門とする
  - 統計解析/データマイニング
    - S-PLUS,Visual Mining Studio等
  - 数理計画法-最適化
    - NUOPT

数理システムへのお問合せは

株式会社 数理システム 営業部
(東京都新宿区新宿2-4-3 フォーシーズンビル10F)

TEL: 03-3358-6681

E-mail: splus-info@msi.co.jp

URL: http://www.msi.co.jp/

- S-PLUS**無料紹介セミナ 定期開催中**
- **システム構築、データ解析に関するご相談も承ります**