BIG Statistics

Dennis Lin
University Distinguished Professor
Department of Supply Chain & Information Systems
The Pennsylvania State University

25 October, 2006

BIG Statistics

- Business Statistics
- Industrial Statistics
- Governmental/Official Statistics (Official Statistics)

Knowledge Discovery in Sciences

- Deduction
  - Physics
  - Mathematics
- Induction
  - Biology
  - Chemistry
- Statistics is more powerful in empirical study & experience accumulation, namely "induction" type.

Business Statistics

Component of Business Research

Management
Behavior
Quantitative

Business Schools

- Accounting/(Statistics)
- Finance
- Marketing
- Management & Organization
- Insurance & Real Estate
- Management Science & Information Systems
- Supply Chain Management
- Business Logistic
- International Business
- Business Administration
- Economics

Business Intelligent

What is the same?

What is new?

What is the Same?

Elements of the exchange process

- A buyer
- A seller
- Buyer and seller can find each other
  (with some way to "authenticate" the other party)
- Each has something of value to offer the other
- They can voluntarily complete the exchange

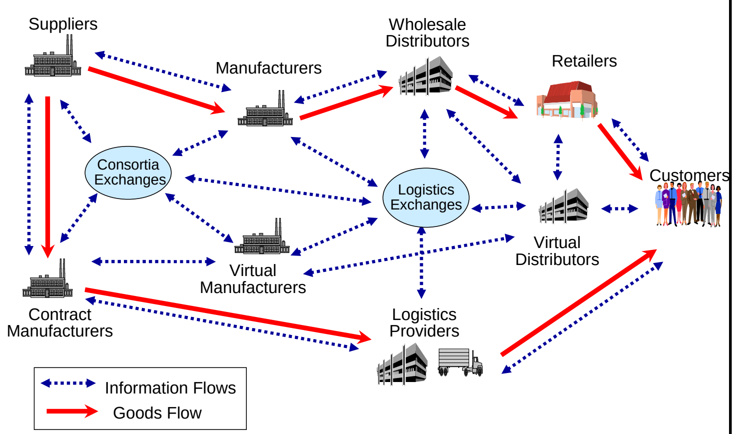
What is New?

- e-business
- Globalization

What is New?

- Technology
- Information System
- e-Business: B2B & B2C
- Supply Chain
- Value Net
- Globalization

What is new? e-Supply Chain



What's different about the electronic exchange process?

1. **Buyers have more control in the exchange process (To some extent, the online medium also helps sellers find buyers)**
2. **Customers behave differently online**
3. **The exchange process for digital products is being completely redefined**
4. **The exchange process for non-digital products is being transformed**
5. **It redefines the way marketing information is gathered, analyzed, and deployed**
6. **The medium alters the nature of competition**

Time Series Analysis

- Time Series Analysis (Univariate)
  - Linear, ARIMA
  - Non-linear
- Time Series Analysis (Multi-variate)
- ARCH & GARCH Model
  - AutoRegressive Conditional Heteroscedasticity
  - GARCH, GARCH-M, Expo-GARCH etc
- Long-Memory, high-frequency data
- Ito's Lemma and Black-Scholes Differential Equation
- Financial Time Series
- Value at Risk

The Newsboy Problem

- An international newsstand must decide how many copies, Q, of the Toronto Star to stock.
  - The owner can purchase papers wholesale at $0.80 each, and sell them for $1.20.
  - Leftover paper are sold to a recycling at $0.05 each.
- **Profit** for the day=

  Min (Q,X) * $1.20

  + Max (0, Q-X) * $0.05

  –Q* $0.80

  where X is the Demand for the day.

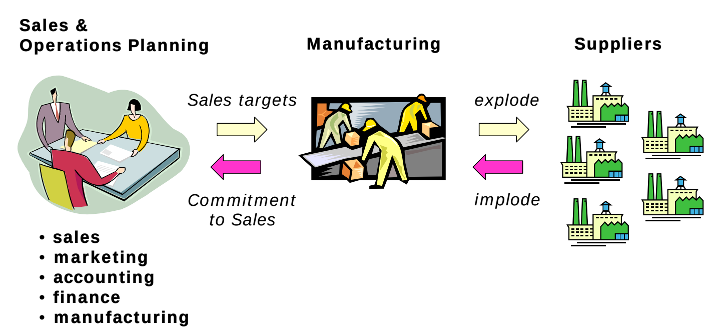
The Newsboy Problem

- Solve for Q which maximize the Profit

Suppose a historical demands are available…

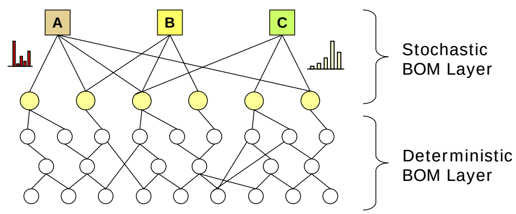
- Optimization problem:
  - X is known (or use x-bar)
- Statistical Problem:
  - X follows Normal distribution (say), or
  - Density Estimation for X
- *Eventually most OR problems can be (more or less) converted to Stat problems*

Configure-To-Order (FTO) System



Unfortunately, Life is Not Simple

- Multiple configurable products
- Component commonality
- Multi-period rolling planning horizon
- Multi-level Bill-of-Materials



Problem Statement

- Given:
  - A set of m products and n components
  - A set of sales targets over T weeks
  - A set of component supply commitments over T weeks
- Determine:
  - A set of commitments to sales over T weeks
- Subject to:
  - Supply constraints
- Objective:
  - To maximum expected profit over T weeks while minimizing the penalty for deviating from the sales targets

Close-Loop Supply Chain

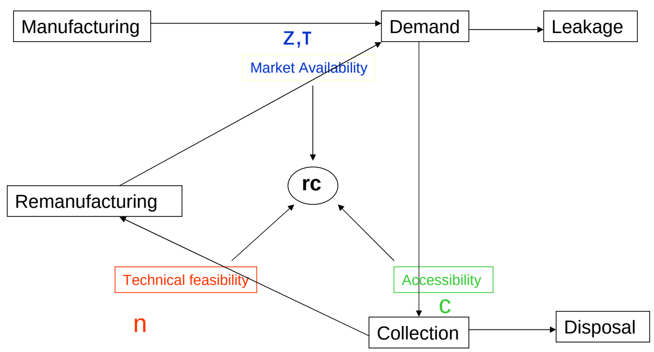
Luk N. Van Wassenhove

Dan Guide (Penn State)

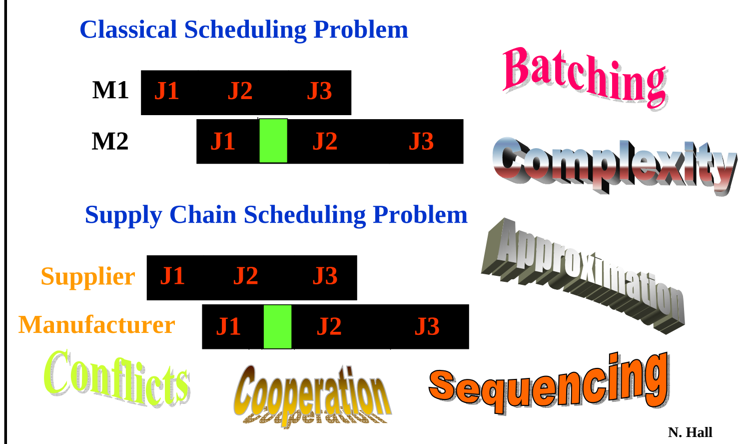
Focus:

- **End-of-use Returns**
- **Perfect Substitution**
- **Value Recover**

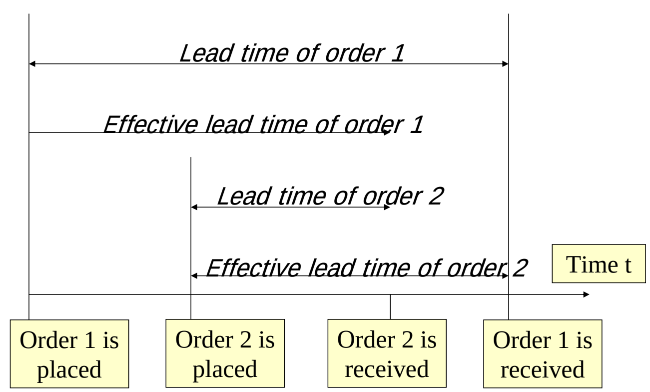
Model :



What is Supply Chain Scheduling?



Inventory: Lead Time



SCHEME

	Demand Rate Deterministic	**Demand Rate Stochastic**
Lead Time Deterministic	***Model 0***	***Model 1***
Lead Time Stochastic	***Model 2***	***Model 3***

Common Assumption: Exponential Distribution

VRP Problem

- Travel Salesman Problem (TSP)
- Chinese Postman Problem (CPP)
- Lawn Mowing Problem (LMP)

- Vehicle Routing Problem (VRP)
  - *m-TSP*

Link Analysis

What are the issues here?

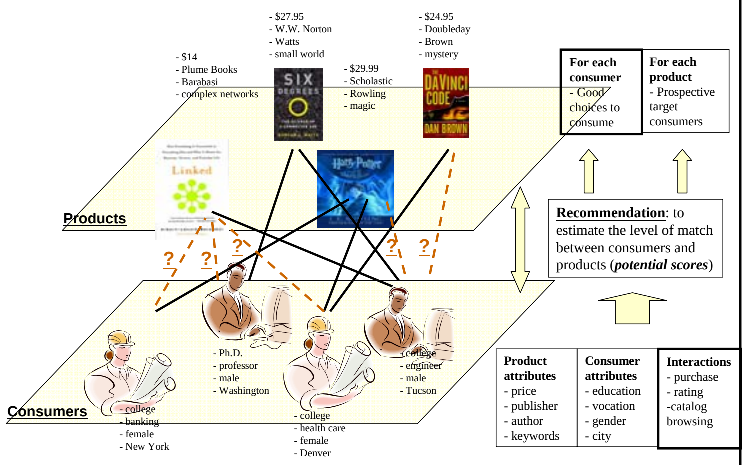
- How do you “quantify” (measure) your observations (people or dog)?
- How do you “characterize” your observations?
- How do you classify (match) them?
- Others?

Recommender Systems

Zan Huang

- Recommender systems
  - Automatically recommend items to users based on product attributes, consumer attributes, and consumers’ implicit or explicit feedback on products [Resnick et al. 1994; Shardanand and Maes 1995; Hill et al. 1995; Resnick and Varian 1997]
  - Products: Discussion postings, webpages, movies, jokes, news, research papers, books, etc.
  - Consumers: Information seekers, online shoppers, etc.

The Recommendation Problem



Topological Measures

- Average path length, L
  - Defined as the average of path lengths of all connected vertex pairs
- Clustering coefficient, C
  - C_Δ [Newman et al. 2001]

$$C_\Delta = \frac{3\times(\text{number of triangles in the graph})}{\text{number of connected triples}}$$

    – A ***connected triple*** is three vertices *x-y-z*, with both vertices *x* and *z* connected with *y* (note that *x-y-z* and *z-y-x* are considered the same connected triple)
- Degree Distribution
  - $P(k)$: The probability for a randomly selected vertex to have degree k
    - The ***degree*** of a vertex is defined as the number of edges connected with it

Theoretical Prediction of Unipartite Graphs Projected from Random Bipartite Graphs

- Average degree

$$z = G_0'(1)$$

- Average path length

$$L = 1 + \frac{\ln(N / G_0'(1))}{\ln\left(\left(\frac{f_0''(1)}{f_0'(1)}\right)\left(\frac{g_0''(1)}{g_0'(1)}\right)\right)}$$

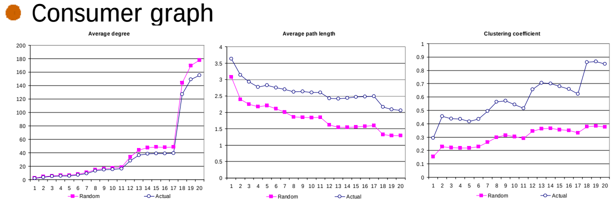
- Triangle clustering coefficient

$$C_\Delta = \frac{M}{N}\frac{g_0'''(1)}{G_0''(1)}$$

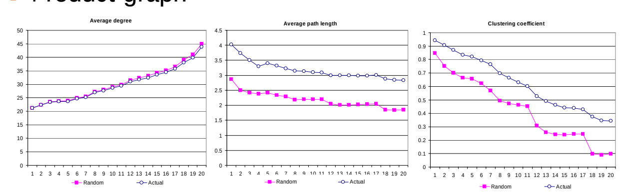
- The predictions for the product graph can be derived similarly by interchanging f and g and interchanging M and N

Actual vs. Random – Book Dataset

- Consumer graph



- Product graph



How to measure/characterize a bipartite graph?

None we know of!!!

Process Mining

case identifier	task identifier
case 1	task A
case 2	task A
case 3	task A
case 3	task B
case 1	task B
case 1	task C
case 2	task C
case 4	task A
case 2	task B
case 2	task D
case 5	task E
case 4	task C
case 1	task D
case 3	task C
case 3	task D
case 4	task B
case 5	task F
case 4	task D

Table 1. A process log.

Process Mining

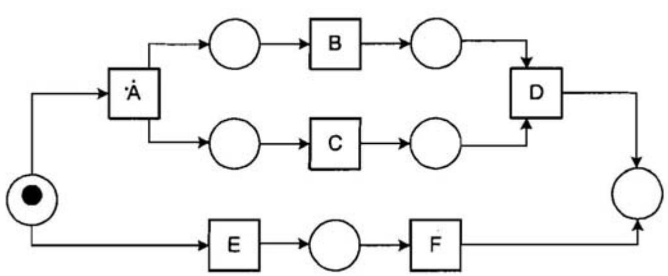


Fig. 1. A process model corresponding to the process log.

Potential Subjects

- Finding General Structure/Network
- Finding systematic pattern
- Identify Unusual Observation (Transaction)
- Model Building

  Data/Table ←→ Model/Graph

Service Performance Analysis and Improvement for a Ticket Queue With Balking Customers

Susan H. Xu
Department of Supply Chain and Information Systems
Penn State University

Characteristics of Ticket Queues

- A newly arriving customer observes:
  - Ticket number
  - Panel display number

    D = ticket number – panel display number

    = ticket position
- Customer's natural tendency is to assume all *D* positions are "real" customers, ignore the possibility that some have balked
- Customer makes joining/balking decision based on *D* (if *D* is too large a customer may balk)
- Let *N* = *D* – number of balking customers

    = queueing position
- *D* is *observable* and *N* is unobservable

Differences Between Ticket Queue & Physical Queue

- ***Physical queue*** provides *complete* information of *N* (i.e., the number of actual customers in system)
  - Customers make joining or balking decision based on *N*
- ***Ticket queue*** provides *incomplete* information of *N*
  - Customers only know *D*, an upper bound of *N*
  - Customers Make joining or balking decision based on *D*

Questions to Be Addressed

- **How to evaluate performance of ticket queue?**
  - What are the distributions of *D* and *N*?
  - What is the balking rate and system utilization?
  - What is customers expected waiting time given his ticket position?
- **How to improve service performance of a ticket queue (reduce the balking rate)?**

Two Specific Example:

RFID & Search Engine

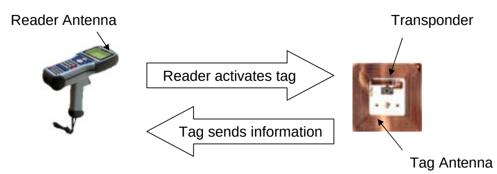
The components of a 96-bit electronic product code (in Hex)

01	0000ABC	000123	000056789
Header 8 bits	EPC Manager (will be assigned to companies) 28 bits	Object Class - Product Code 24 bits	Serial Number 36 bits

The RFID tag responds to the reader by broadcasting its EPC, which is a 96-bit code consisting of:
· 8 bits of header information
· 28 bits identifying the organization that assigned the code
· 24 bits identifying the type of product
· 36 bits representing serialization information for the product

Source: Avicon white paper, 2003

RFID



• Radio Frequency Identification (RFID) technology has been in use since the 1950's

•Advocated as Electronic Product Code™ or EPC™

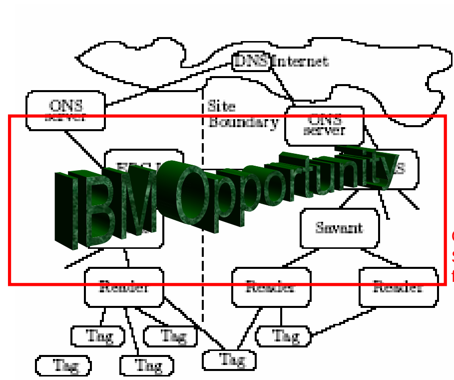
RF Communication

- Electromagnetic waves modulated to carry data/signals
- Two different ways to generate ways
  - Inductive coupling
    - Close proximity electromagnetic wave
  - Propagating electromagnetic waves
- The fundamental RF communication theories apply—nothing new.
- New: the cost, size, signal processing capability.

Impact to Statistics

How to analyze the population data?

Architecture



ONS: object name server
Maps EPC→ URL

EPC Information service
Higher level service for apps

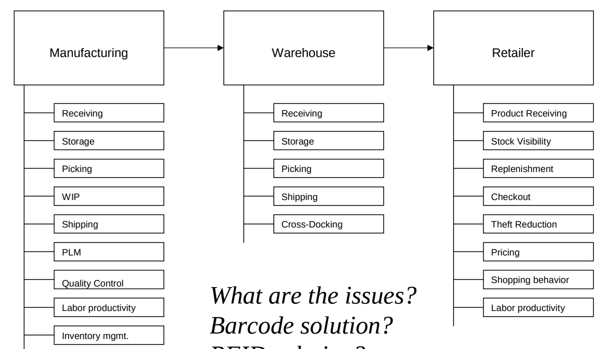
Gather data from readers:
Smoothing, coordination,
forwarding,etc.

Source: Chawathe, et al, VLDB Conference proceedings, 2004

RFID vs. Barcode

Lin, Dennis K.J. and Wadhwa, Vijay

Basic Structure



What are the issues?
Barcode solution?
RFID solution?
Comparisons!!!

S.I.T. Space

- State
- ID
- TimeStamp

- Particularly interested in the difference between *Execution* & *Planning*
  (Sense & Response)

An IBM RFID Example on Analysis:

What you're looking for in sensor data?

Lin, Shu and Wadhwa

Preliminary Analysis on Wal-Mart RFID data

Introduction

- 5 products (Lemonade, Gator Fruit Punch, Jumbo CapNCrunch, Oat variety and Gator Orange).
- Data gathered from Dec 05 to June 06 in 274 stores.
- 9 unique locations where readers are located.

Location-Scan Information

Location	Count of Read
100	7088
101	25
102	55
103	710
104	33
105	15678
112	20878
117	313
125	2

Data Information.

- There are 35,228 distinct EPC's which are scanned. Of these 35,228 EPC's 7918 belong to Gator orange, 3860 to Oat Variety, 6685 to Jumbo CapNCrunch, 9156 to Gator Fruit Punch and 7609 to Lemonade.
- Each EPC is scanned anywhere between 1 time to 2000 times. This could be attributed to many factors like a pallet being left near box-crasher (location 105).
- 9617 EPC's belong to more than one store. In this case an EPC can belong to anywhere from 2 to 37 stores.

Flow-Analysis

- 3 Benchmarked process flow:
  - Retail Store: 100,112,112,105
  - DC with Conveyor: 100,103,101
  - DC with Shrink-wrap:100,117,101
- Objective: Compare the actual process flow to benchmarked process and find if possible the reasons for deviations.
  - Flow-analysis or the flow time can be related to the product demand.
- Assumption: In the first stage only those EPC's which belong to one store are analyzed.

Average Daily Demand

- The average daily demand can be computed based on the given data which gives an indication of the product consumption.
  - Here it is assumed that each unique EPC represents one unit of demand.
  - The average daily demand of Gator Orange is 44 units.
  - The average daily demand of Oat variety is 21 units.

Future Directions & Issues

- Need to understand how the data is gathered and the process behind it.
- Efficiency at various stores/DC can be computed using the time lags between the RFID scans.
- Infer why the actual process differs from the benchmarked process.

Search Engine & Citation Index

Google's Page Rank formula

- The PageRank of a page A is given as follows:

$$PR_1(A) = (1-d) + d \times \left(\frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \ldots + \frac{PR(T_n)}{C(T_n)} \right)$$

❖PR(A) is the PageRank of page A;

❖ PR(Ti) is the PageRank of pages Ti which link to page A;

❖ C(Ti) is the number of outbound links on page Ti;

❖ d is a damping factor which can be set between 0 and 1; usually set to **0.85**

❖ n is the total number of all pages which link to page A.

Markov Chains

- Matix A

$$a_{ij} = \frac{(1-d)}{N} + d\frac{g_{ij}}{c_j} \qquad d=0.85$$

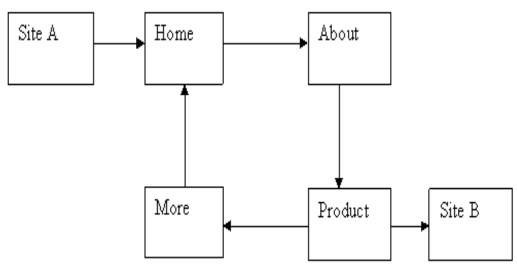
- Matix A max eignvalue =1

$$Ax=x \qquad \sum_i x_i = 1$$

- Matix A eignvector = PageRank(k)

$$x_k = \sum_{j=1}^{N} a_{kj} x_j = \frac{(1-d)}{N} + d \sum_{g_{kj}=1} \frac{x_j}{c_j}$$

Example 5



Example 5

$$H = (1-d) + d(\frac{M}{1} + \frac{SA}{1})$$

$$A = (1-d) + d(\frac{H}{1})$$

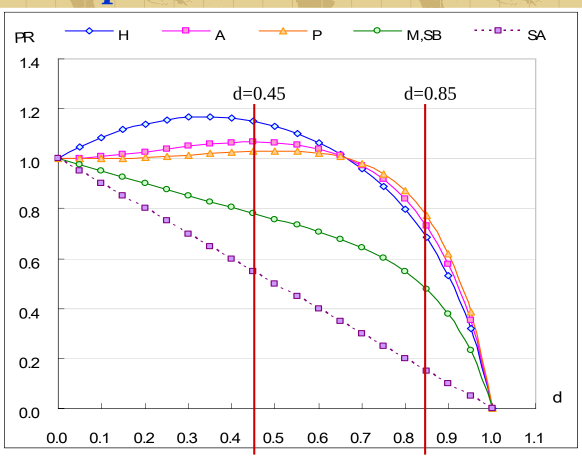
$$P = (1-d) + d(\frac{A}{1})$$

$$M = (1-d) + d(\frac{P}{2})$$

$$SA = (1-d)$$

$$SB = (1-d) + d(\frac{P}{2})$$

Example 5

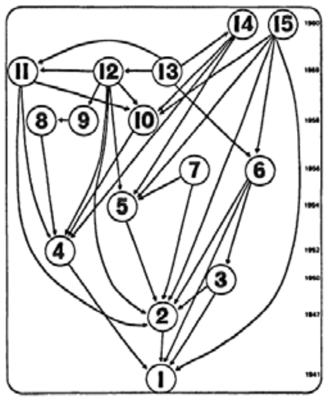


How to Increase your PageRank?

How to Increase Your Paper Citation?

- Individual article
- Journal
- How is this (Impact Factor) related to PageRank?

Paper Citation: Garfiled (1972)

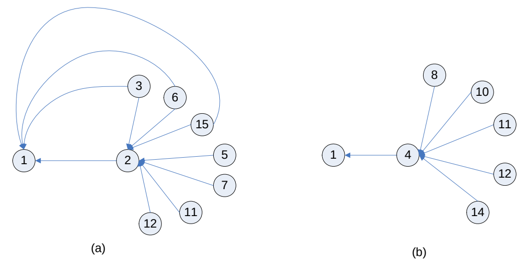


Number of Citations:

Paper#1: 5
Paper#2: 7
Paper#4: 5

Garfiled (*Science*, 1972)

Paper Citation: Citation Analysis



•Paper#1 is important
(than Paper#2 which is likely to be an extension or review)
•Paper#4 is equally important to Paper#1
•Paper#4 is more important than Paper#2

Few Remarks About Citations

- The median of the number of citations, across articles from all disciplines, is 0.
- For an "improved" measurement on "Citation Analysis,"
  see Lin *et. al.* (2006).

Industrial Statistics

Industrial Statistics *(Lin)*

- Process Capacity Index (PCI)
  - Chao and Lin (2005)
- Control Chart
  - Yeh and Lin (2005)
- Reliability
  - Highly reliable products
- Design of Experiment
  - Lin's recent work (talk at Australia)
- Quality Assurance & Six-Sigma Development
- Data Management & Value Net
- Relational Learning & Industrial Classification
- Others

Process Capacity Index

- C_p
- C_{pk}
- C_{pm}
- C_{pmk}
- $C_p(u,v)$
- C_{pw}
- etc.

Basic Idea:

Process yield **= 2Φ(3*C_y)-1**

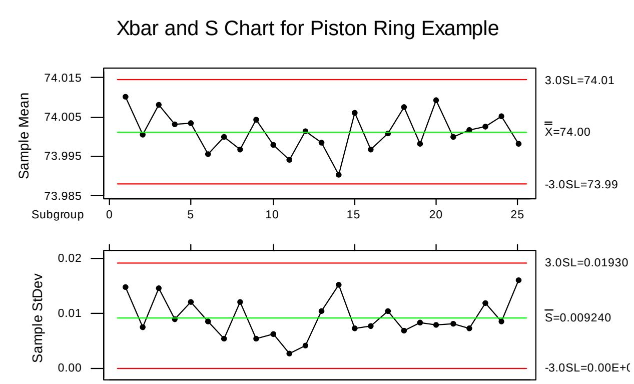
Thus, we (Chao and Lin, 2005) have

$$C_y = \frac{1}{3}\Phi^{-1}[\frac{1}{2}(F(USL) - F(LSL) + 1)]$$

$$\hat{C}_y = \frac{1}{3}\Phi^{-1}[\frac{1}{2}(F(USL;\hat{\theta}) - F(LSL;\hat{\theta}) + 1)].$$

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n}\Phi(\frac{x - X_i}{1.06sn^{-1/5}})$$

Shewhart Control Charts

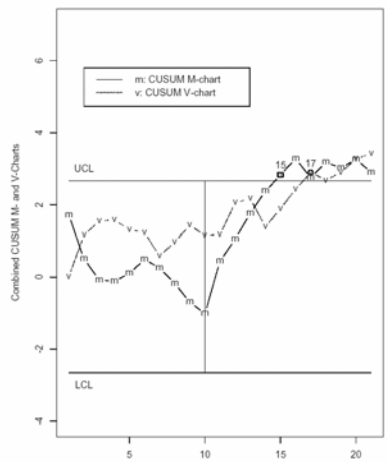


$$m_j = P\left(t_{N-k} \leq \frac{\overline{X}_j - \overline{\overline{X}}}{\overline{S}\sqrt{\frac{1}{n}+\frac{1}{N}}}\right)$$

$$v_j = P\left(F_{n-1,N-k} \leq \frac{S_j^2}{\overline{S}^2}\right).$$

$$S_m(t) = \sum_{j=1}^{t}(m_j - \frac{1}{2}) = (m_t - \frac{1}{2}) + S_m(t-1) \text{ and}$$

$$S_v(t) = \sum_{j=1}^{t}(v_j - \frac{1}{2}) = (v_t - \frac{1}{2}) + S_v(t-1),$$



Yeh and Lin (2003)

MULTIVARIATE CONTROL CHARTS FORMONITORING COVARIANCE MATRIX: A REVIEW

Quality Technology & Quality Management

Yeh, A.B., Lin, Dennis K.J. and McGrath R.N. (2005)

Problem Description

$X = (X_1, X_2, ..., X_p)'$, p correlated quality characteristics

$X \Leftrightarrow N_p(\mu, \Sigma)$

When the process is in control, $\mu = \mu_0, \Sigma = \Sigma_0$

Question:

How to Devise a Control Scheme to Detect Changes in

Σ (e.g., to $\Sigma_1 \neq \Sigma_0$)

Important Considerations

- The Methodology a Control Chart Uses to Combine Sampled Information – Shewhart, CUSUM and EWMA
- Subgroup Size: $n > p$, $n = 1$, $1 < n < p$
- What types of Changes in Σ is the chart designed to detect?

Recent Research Focus: Control Charts

- Multivariate Control Charts
- Run-to-Run Control Charts
- Control Charts for Data-Rich Environment
- Cause-Selecting Control Charts
- Control Chart for Profile Data
- Control Chart for Functional Response

Design of Experiment

How to collect useful information?

Design of Experiment (Lin)

- Multiple Response Problems
  - Optimization: Kim and Lin (*JRSS-C*, 2000)
  - Design: Chang, Lo, Lin & Young (*JSPI*, 2001)
- Computer Experiment
  - Beattie and Lin (1998)
- Dispersion Effect
  - McGrath and Lin (*Technometrics*, 2002)
- Foldover Plan
  - Li and Lin (*Technometrics*, 2003)
- Supersaturated Designs
  - Lin (*Technometrics*, 1993, 1995, 2001) and others
- Uniform Designs
  - Fang, Lin, Winker & Yang (*Technometrics*, 1999)

What's Next?

Design for Quality
in
Pharmaceutical Industry

Supersaturated Designs

Lin (UTK Technical Report, 1991)
Lin (*Technometrics*, 1993, 1995, 2001)
and many others

SUPERSATURATED DESIGN

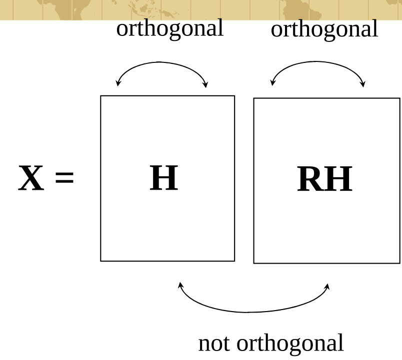
How can we study k parameters with n(<k) observations (experiments)?

A situation for using supersaturated design:

- *A Small number of run is desired*
- *The number of potential factors is large*
- *Only a few active factors*

Supersaturated Design From Hadamard Matrix of Order 12
(Using 11 as the branching column)

Run	Factors											
No.	I	1	2	3	4	5	6	7	8	9	10	(11)
1	+	+	+	-	+	+	+	-	-	-	+	-
2	+	+	-	+	+	+	-	-	-	+	-	+
3	+	-	+	+	+	-	-	-	+	-	+	+
4	+	+	+	+	-	-	-	+	-	+	+	-
5	+	+	+	-	-	-	+	-	+	+	-	+
6	+	+	-	-	-	+	-	+	+	-	+	+
7	+	-	-	-	+	-	+	+	-	+	+	+
8	+	-	-	+	-	+	+	-	+	+	+	-
9	+	-	+	-	+	+	-	+	+	+	-	-
10	+	+	-	+	+	-	+	+	+	-	-	-
11	+	-	+	+	-	+	+	+	-	-	-	+
12	+	-	-	-	-	-	-	-	-	-	-	-



Thus Permute rows of RH to minimize $E(s^2)$, say.

UD OD SSD

$$U \oplus L = \begin{bmatrix} 1 & 1 \\ 2 & 7 \\ 3 & 3 \\ 4 & 9 \\ 5 & 5 \\ 6 & 6 \\ 7 & 2 \\ 8 & 8 \\ 9 & 4 \end{bmatrix} \oplus \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 2 & 2 & 2 \\ 1 & 0 & 1 & 2 \\ 1 & 1 & 2 & 0 \\ 1 & 2 & 0 & 1 \\ 2 & 0 & 2 & 1 \\ 2 & 1 & 0 & 2 \\ 2 & 2 & 1 & 0 \end{bmatrix} = X = \left[\begin{array}{cccc|cccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 2 & 0 & 2 & 1 \\ 0 & 2 & 2 & 2 & 0 & 2 & 2 & 2 \\ 1 & 0 & 1 & 2 & 2 & 2 & 1 & 0 \\ 1 & 1 & 2 & 0 & 1 & 1 & 2 & 0 \\ 1 & 2 & 0 & 1 & 1 & 2 & 0 & 1 \\ 2 & 0 & 2 & 1 & 0 & 1 & 1 & 1 \\ 2 & 1 & 0 & 2 & 2 & 1 & 0 & 2 \\ 2 & 2 & 1 & 0 & 1 & 0 & 1 & 2 \end{array}\right]$$

Fang, Lin & Ma (2000)

SSD: Looking Ahead

Supersaturated Design

- SSD is much more mature than ever.
- Micro-Array Design and Analysis
- Computer Experiment: Model Building (using SSD)
- Higher Level SSD
- Spotlight Interaction Effects (Lin, 1998, QE)
- Combination Designs: Rotated FFD & SSD

Computer Experiment

Beattie and Lin (2005)
How to run expensive simulation?

Where have all the Data Gone?

- No need for data (Theoretical Development)
- Survey Sampling and Design of Experiment (Physical data collection)
- Computer Simulation (Experiment)
  - Statistical Simulation (Random Number Generation)
  - Engineering Simulation
- Data from Internet
  - On-line auction
  - Search Engine

Statistics vs. Engineering Models

$$y = f(x,\theta) + \varepsilon$$

Statistical Model

$$y = \beta_0 + \Sigma\beta_i x_i + \Sigma\beta_{ij} x_i x_j + \varepsilon$$

Two-Cents on Random Number Generation

- Random Number Generator
  - Deng and Lin (2000, *The American Statistician*)
- Transformation to Non-Uniform Distribution

Goals—Computer Experiment

- Confirmation
- Sensitivity Analysis
- Empirical Model Building
- Optimization
- Model Validation
- High Dimension Integration

A Structured Roadmap for Verification and Validation--Highlighting the Critical Role of Experiment Design

James J. Filliben

National Institute of Standards and Technology
Information Technology Division
Statistical Engineering Division

2004 Workshop on Verification & Validation of Computer Models of High-Consequence Engineering Systems
NIST Administration Building
Lecture Room D
3:10-3:25, November 8, 2004

Reliability

More on Highly reliable Products

Run-to-Run Problems

Response Surface Methodology: 50 Years Later

Dennis K.J. Lin
The Pennsylvania State University
DennisLin@psu.edu

Ambitious Goal

- What is Response Surface Methodology?
- What type of problems they had in mind back to 1950?
- What was available in 1950?
- What type of problems today (50 years later)?
- What is available today?
- Can we do something significantly different?

Quality Assurance & Six-Sigma

Total Quality Management (TQM)

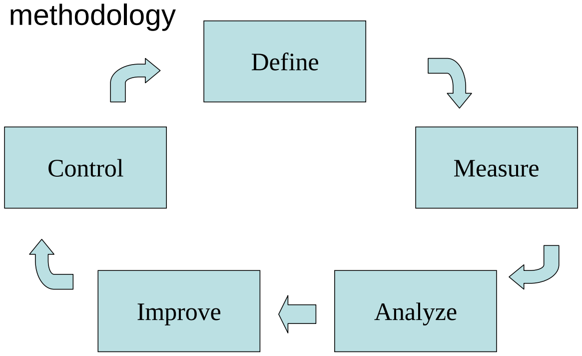
Taguchi Method

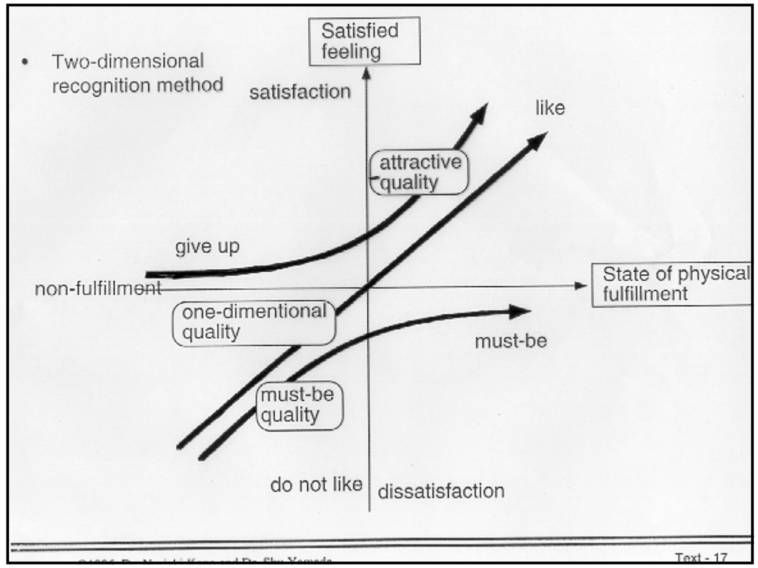
Six Sigma ...

Green Lean Six Sigma

Six Sigma Project Process

- Rigorous, data-driven and customer-focused methodology





What's Next?

Quality in
Service Industry
(Service Science)

Governmental Statistics

Official Statistics

Official Statistics

- 強國需知十三數

--商鞅 (390 B.C.)

- Index Management
- Sampling

Index Management

- What to measure?
- How to measure?

- 人 (population):
  Age, geographical, etc and distribution
- 錢 (Wealth): 國富調查

- Economics Index
- Social Index

Government/Official Statistics

- Index Management
- From Census to Sampling
- Sampling using modern technology
  (Web, Palm)
- Error in Variables
- Missing values
- Data Mining

Send $500 to

- Dennis Lin

  University Distinguished Professor

  483 Business Building

  Department of Supply Chain & Information Systems

  Penn State University

- +1 814 865-0377 (phone)
- +1 814 863-7076 (fax)
- DKL5@psu.edu

(Customer Satisfaction or your money back!)