

タグチメソッドとは・リローデッド

非常に重要な工学手法なのです。一刻もはや手法を活用しなければ他社に負けてしまいます！

そもそもタグチメソッドとは何でしょうか(^^)?

タグチメソッドは田口玄一博士が提唱された手法で、米国の自動車業界を蘇らせた「魔法の杖」と呼ばれています。

タグチメソッドには、次のような特徴があります。

1. 設計 開発段階で適用する工学手法です。
2. 自然の法則に逆らって、自然界にないものを創作するのが工学です。実験計画法は科学です。タグチメソッドは実験計画法と違います。
3. ばらつきを抑える、つまり安定性を確保するのです。ばらつきを抑えてから目標値にもっていくのです。
4. ばらつきの指標に「SN比」というものを使います。
5. 誤差因子、制御因子、信号因子という因子(要因)を考えます。
誤差因子」
制御因子」
信号因子」
6. L18直交表にしたがって実験を計画します。
7. 実験結果からSN比を計算します。
8. 最適条件をみつけ、現行条件に比べどれだけばらつきがすくなくなったか、つまり安定したかを評価します。
9. 最適条件で実際に確認実験をします。

ランダムですが、先日のセミナーでポイント注意点を受講者に紹介しました。タグチメソッドのポイントを思いつくまに書きました。

A社製とB社製の部品の評価はSN比を用いれば良い！

広瀬・土田編著『Excelでできるタグチメソッド解析法入門』同友館 刊の99ページ参照。

SN比を使わないとしたら何を使う？.....それは、「ヒストグラム」！

L18直交表を使おう！

何故か？？18回の実験で8要因の効果を調べることができるからです。主効果だけ考えることができ、交互作用は列に現われないようになっている。

L18直交表を使用しないと実験回数は $2 \times 3^7 = 4374$ 回になる。L18直交表だと実験回数は18回で済みます。

外側因子(誤差因子 信号因子は使用する側の条件)、内側因子(制御因子)は設計する側の条件。

L18直交表(2水準と3水準が混在する)を使う。しかし4水準、6水準有るときはどうするか?!..... L18直交表で1, 2, 3, 4, 1, 2, 3, 4,と新しい列を作れば良い。ただし直交性は崩れます。

L18直交表で5因子しかないときは、5つの列だけを使えば良いのです。(詳細は学習してください。)

L36直交表は3水準の因子を13個まで割付けることができる。

説明変数にすると26。Excelの回帰分析は説明変数の最大は16。さあどうする？ Excelのソルバーを使おう！

(Excelでできる最適化の実践らくらく読本』同友館 刊 参照)

影響度は最大値から最小値を引き算した値。したがって必ず正の値になります。

はじめに、誤差因子だけL12直交表を用いて取り出すのも一案。

実験計画法は科学だ！(真実の発見 因果分析) タグチメソッドは工学(自然界に無いものを創造する)

SN比は加算 減算ができる。それはlogを用いているから。減算したものが利得となります。

回帰分析法でタグチメソッドデータを解析しているので、例えばL18の18番目のデータ(y18)が無くても解析できる(精度は落ちます)。変動分解法でも可能だが計算が面倒(-;-)。

望大特性、望小特性とも望目特性に帰着できる。

したがって、静特性では望目特性が重要。動特性は望目特性をダイナミック(動的)に動かしたもの。したがって、静特性は動特性に含まれる。

タグチメソッドは直交表L18 L36、制御因子 誤差因子(著者らは環境因子と呼び変えている)、信号因子、SN比のアイデアの組合わせである。

L12(わずか12回の実験で11要因まで調べることが可能)直交表で要因の効果の有無を調べる(言わば「リトマス試験紙」)、L18 L36で最適条件まで求める。

マーケティング分野版実験計画法であるコンジョイント分析は望大特性版タグチメソッドである。
タグチメソッドは設計者のアイデアの良さ・悪さをはやく評価(わかる)する手法。うまくはやくやすく楽しく行う工学手法

タグチメソッドを適用するとき、一番重要な点は基本機能。この基本機能を示すグラフが夢に出るようになるう(笑)！ <http://www.datamining.jp/ma/20031119.htm>

タグチメソッドの詳細は広瀬・上田編著

Excelでできるタグチメソッド解析法入門 』同友館あるいは弊社の通信教育講座で学習して下さい。

<http://www.datamining.jp/tsushin/h-taguchi3.htm>

セミナーの開講講座リサーチについて(お願い)

来る2004年2月の平日(1日間)に、東京で上田データマイニング塾主催でデータマイニングセミナーを開催に向けて準備中です。そこで、今回は皆さんのご意見も採り入れてみようかなと思ひ、こんな形でリサーチをすることにしました。ズバリ! あなたはどの講座を受講したいですか(??)以下の要領をお読みになってお答え下さい。

回答は、2クリックの簡単なものです。

回答はこちらから! <http://www.datamining.jp/cgi-bin/enq.cgi>

各講座の詳細は、こちらをご覧ください。

http://www.datamining.jp/seminar/h-semi_curriculum.htm

上田塾長が講師を務める公開セミナー続々開催!

- ・12月2日(火)Excel徹底活用 多変量解析
- ・12月8日(月)初めて学ぶ・エクセルでのデータマイニング
- ・12月11日(木)アンケートの設計とデータ分析 活用のすすめ方

<http://www.datamining.jp/seminar/h-seminar.htm>

通信教育講座 好評開講中!

このような予測を行う場合でも、Excelによる簡単な操作で、誰でも同じように予測を行うことができます。

<http://www.datamining.jp/tsushin/t-yosoku/suuryouka1.htm>

日によって、売上高の増減に影響を及ぼす項目(要因)が曜日や天候などの数値以外のデータであっても、簡単に予測を行う例を挙げています。通信教育講座「予測入門」では、このように事例紹介形式で、Excelの操作方法を含め、わかりやすく解説しています。また、「予測入門」をご受講にあたり、「データマイニング入門」を受講されると、更に理解が深まります。 <http://www.datamining.jp/tsushin/h-nyumon1.htm>

「データの標準化」について・1

データ解析の前に、「データの標準化」を考えることが重要になります。わかりやすい例で説明をすると、例えば身長・体重・視力・虫歯の本数などの身体データの標準化をします。

身長は、例えば「170cm」、「183cm」と表され、体重は「48kg」、「60kg」と表されますが、では「虫歯の本数」や「視力」はどうでしょうか。……「視力」は「0.1」、「1.5」……、虫歯は「1本」、「2本」と表されますね。このように、単位(の種類)が混在するデータを同時に解析する場合、標準化が必要な場合があります。同じ「150」という数値でも「(身長)150cm」と「(体重)150kg」は等しくなく関連もありません。しかし、一般的には、身長の割には体重が多いなど、身長と体重を比較して考えることがあります。このように、それぞれの数値を一定の尺度に置き換える作業のことを「標準化(または基準化)」と呼びます。

ここで標準化の方法の1つとして、手順を簡単に説明しましょう。

【標準化の方法】

- 1) No. 1~No. 10までの10個のデータがあるとします。
この10個のデータの平均値(=AVERAGE)を求めます。
- 2) 10個のデータの標準偏差を求めます。(=STDEV)
- 3) まずNo. 1のデータについて、次の計算をして、標準化させた値を求めます。
([No.1の値]-[10個の平均])÷[10個の標準偏差]
- 4) 3)の計算をNo. 2からNo. 10についても行います。
- 5) 全ての標準化させた値の平均値は0になり、どの項目においても、同じレベルの数値になりました。

「AIC (Akaike's Information Criterion)」

~~ 予測問題と解決 連載 第8回 統計モデル どちらのモデルが良いか ~ 統計ではモデルを考えます。モデルといってもファッションモデルのモデルではありません。統計モデルのことです。たとえば、M1とM2の回帰モデルM1: $y = a_0 + a_1x_1 + a_2x_2 + a_3x_3$ M2: $y = b_0 + b_1x_1 + b_2x_2$ を考えます。(M1、M2それぞれ、 a_0 b_0 は切片を表します。) M1はyを被説明変数、説明変数が x_1 、 x_2 、 x_3 の回帰モデル、M2は説明変数が x_1 、 x_2 の回帰モデルです。どちらのモデル

ルが良いかを、どのように判定すれば良いのでしょうか。元 統計数理研究所長の赤池 弘次博士はモデルの良さ(悪さ)を計る情報量規準を提案しました。世界的に有名な「AIC (Akaike's Information Criterion)」です。回帰モデルのとき、AICは次のようになります。AIC=データ数* $\log(1 - \text{重相関係数の自乗}) + 2 * \text{説明変数の個数}$ です。ここで \log は自然対数です。

M1とM2のAICを求め、小さい方を良いモデルとします。たとえば、M1のAICが-3.50で、M2のそれが-4.23ならば、モデルM2の方が良いモデルと判定するわけです。回帰モデルが3つ以上あっても同様です。AICが最小のものを最適なモデルとします。AICは説明変数選択規準のひとつです。日経の「e-ラーニングのデータマイニングの活用編」で出てくるRuも説明変数選択規準です。Ruは大きい方が最適な回帰モデルとなります。Ruで最適な回帰モデルを選んでも、AICで選んでも(ほぼ)同じ結果になります。

データマイニング

データマイニングとは通常膨大なデータをマイニング(採掘)して宝物(情報、知見、知識、課題、仮説など)を見つける手法・プロセスのこと。代表的な手法には統計、ニューラルネット、パターン認識などがある。データマイニングツールのことをソフトウェアと呼んでいる。

データウェアハウス

データの倉庫のこと。通常は膨大な(ギガバイト、テラバイト級)データを入れる。データマイニングとデータウェアハウスはペアで使われる。著名な商品にEssBase、RedBrick、diaprisimなどがある。

外れ値

数値データのなかで、極端に大きな、あるいは小さな値をとるデータのこと。異常値、例外値、特異点ともいう。ソフトウェアを用いて外れ値をみつけ、原因を追求することで知見が得られることがある。データマイニングらしい手法の1つ。

代表値

代表値には、平均、中央値、最頻値などがある。

平均

平均には単純平均(相加平均)、幾何平均、調和平均などがある。単純平均はデータを加えてデータの個数で割った値で有る。単純平均>幾何平均>調和平均となる性質がある。

中央値(メデアン)

データを小さい順にならべたとき、丁度真ん中(中央)にある値のこと。データが偶数個のときは、中央の2つの平均値が中央値

最頻値(モード)

最頻値とはファッションモードと同じように、最も多い数字のこと。4,6,8,8,8,9,9,10,11,13の最頻値は8。

ばらつき

ばらつきを表わす統計量には、標準偏差、レンジ(範囲)がある。

標準偏差

標準偏差はばらつきを表わす統計量である。

レンジ(範囲)

ばらつきを表わす統計量はらつき最大値-最小値のこと。

散布図

対になったデータを横軸、縦軸で平面上にプロットしたグラフで、データの様子を捉えるには極めて有効である。対になったデータを解析するには、まず、散布図を描くとよい。外れ値などを見つけることができる。

相関係数 r

ある量とある量との線形な関係度を表わす指標で-1と1の間の値をとる。1に近いときは強い相関があるといいい-1に近いときは負の強い相関があるという。単回帰式の良さの指標でもある。

相関の有無の判定

相関があるかを判定するには、 r の自乗 $> 4 / \text{データ数} + 2$ が成立すれば、相関があるとする。

単回帰式

$y = a + bX$ の式のことAをY切片、Bを傾きあるいは回帰係数という。最小自乗法を用いてABを求める。XでYを表わす式である。

重回帰式

$y = a + b_1X_1 + b_2X_2 + \dots + b_kX_k$ のこと。aをy切片、 b_1, b_2, \dots, b_k を回帰係数という。最小自乗法を用いてa, b_1, \dots, b_k を求める。 X_1, X_2, \dots, X_k を説明変数、yを被説明変数という。複数の説明変数でyを表わす式である。回帰式、回帰モデルともいう。

重相関係数 R

重回帰式のよさを示す指標で0と1の間の値をとる。一般には1に近いほど良い回帰式である。

最適な回帰モデル

説明変数のなかで、本当に y に対して効いているもので回帰モデルを構成することがポイントである。ケチの原理といっているだけ少ない説明変数で回帰モデルはつくるべきである。

最適な回帰モデル作成法

EXCEL の分析ツールを用いて最適な回帰モデルを求めるには説明変数減少法による。減らす説明変数は P- 値 (危険率) が最大なものとする。モデル候補のなかで説明変数選択基準 RU が最大のを最適なモデルとする。

説明変数選択基準

最適な回帰モデルを選択するための規準。赤池の情報量規準 (AIC) 芦賀他の規準、佐和の規準、竹内の規準、上田の規準、MALLOW'S の C P などがある。ここでは上田の規準 RU を紹介すると $RU = 1 - 1 - R^2 \times (\text{データ数} + \text{説明変数の個数} + 1) / (\text{データ数} - \text{説明変数の個数} - 1)$ である。

ダミー変数

好き嫌い男女など定性的情報を 0, 1 データであらわして、回帰式を求めることができる。この 0, 1 データのことをダミー変数という。数量化理論 類はダミー変数を用いた回帰モデルである。

数量化理論 1 類

林知己夫博士が提案した統計手法。ダミー変数を用いた回帰モデルであることがわかっている Y (回帰モデルの被説明変数のこと) を外的基準という。外的基準をアイテム・カテゴリデータ (ダミー変数のこと) で表現し、回帰係数に相当するカテゴリスコアを求める。カテゴリスコアは定性的な情報を数量化したもの。

数量化理論 類

林知己夫博士が提案した統計手法。外的規準が定性的で 2 グループのときはダミー変数を用いた回帰モデルであることがわかっている Y (回帰モデルの被説明変数のこと) を外的基準という。外的基準をアイテム・カテゴリデータ (ダミー変数のこと) で表現し、回帰係数に相当するカテゴリスコアを求める。カテゴリスコアは定性的な情報を数量化したもの。

判別分析

回帰モデルで y が定性的な情報 (良品、不良品、合格、不合格) などのとき、判別分析モデルという。グループ数が 2 のとき、回帰モデルで解析可能である。

ステレオグラム

縦棒クラフのこと。クロス表をグラフ化するにはステレオグラムがよい。

クロス表

分割法ともいう。事例はテキスト参照。属性の数により 2 重クロス表、3 重クロス表、... がある。EXCEL でクロス表を求めるにはピボットテーブルを使う。複雑なクロス表の分析には双対尺度法 (対応分析) が有効である。

補正 R2

自由と修正重相関係数と呼ぶ。

数量化理論 3 類

コレスポンデンス分析、双対尺度法、主成分分析よりおもしろい結果が得られる。

F 分布

数理統計学で仮説検定に用いられる分布の一つ。一つの正規分布母集団からランダムに抽出した 2 組の試料の不偏分散 (分散) の比がこの分布に従う。分散の差の検定、分散分析法等に利用 (F 検定)。F は R.A. フィッシャー

²分布

数理統計学で仮説検定に用いられる分布の一つ。母集団に m 種のもものがそれぞれ p_1, p_2, \dots, p_m の比率で含まれているとき、これからランダムに n 個の試料をとり出したとき np_1, np_2, \dots, np_m に分布することが期待されるが、実現した試料分布が n_1, n_2, \dots, n_m だったとすると、

(式 1)

の値は m が十分大きいとき、自由度 $m - 1$ の ²分布に従う。出現率や独立性の検定、分散区間推定など用途が広く、これを用いた検定を ²検定という。F 分布 / t 分布 / 正規分布