◆連　載◆

《連載 全 3 回》

第 2 回　R & D 部門におけるデータ共有，利活用のためのデータの記録，蓄積，分析方法

上島 豊　　（株）キャトルアイ・サイエンス　代表取締役

《PROFILE》

略歴：

1992 年 3 月　大阪大学工学部　原子力工学科　卒業
1997 年 3 月　大阪大学大学院工学研究科　電磁エネルギー工学専攻　博士課程修了
1997 年 4 月　日本原子力研究所　博士研究員
2000 年 4 月　日本原子力研究所　研究職員
2006 年 3 月　日本原子力研究開発機構（旧日本原子力研究所）　退職
2006 年 4 月　キャトルアイ・サイエンス設立　代表取締役　就任

主な参加国家プロジェクト：

文部科学省 e-Japan プロジェクト　「ITBL プロジェクト」,「バイオグリッドプロジェクト」
総務省 JGN プロジェクト　「JGN を使った遠隔分散環境構築」
文部科学省リーディングプロジェクト　「生体細胞機能シミュレーション」

主な受賞歴：

1999 年 6 月　日本原子力研究所　有功賞
「高並列計算機を用いたギガ粒子シミュレーションコードの開発」
2003 年 4 月　第 7 回サイエンス展示・実験ショーアイデアコンテスト文部科学大臣賞
「光速の世界へご招待」
2004 年 12 月　第 1 回理研ベンチマークコンテスト　無差別部門　優勝

主な著作：

培風館「PSE book ―シミュレーション科学における問題解決のための環境（基礎編）」ISBN：456301558X
培風館「PSE book ―シミュレーション科学における問題解決のための環境（応用編）」ISBN：4563015598
培風館『ペタフロップス コンピューティング』ISBN978-4-563-01571-8
臨川書店『視覚とマンガ表現』ISBN978-4-653-04012-5

4　複数工程実験のデータは，どのような項目名で記録すべきか？

前章では，「記録するデータの項目名は，意味が分かれば何でも良いという訳ではない」と題し，記録するデータの項目名をどのように決めるべきか説明した。本章では，「複数工程実験のデータは，どのような項目名で記録すべきか？」と題し，複数工程実験のデータの項目名はどのように決めるべきかを論じる。

工程内に関しては，「項目名－項目値」に揃えるのは簡単なのだが，工程を跨いだ時に「項目名－項目値」を維持するのが難しい。実際，「項目名－項目値」の形が維持できなければ，工程を跨いだデータの分析が非常に困難になる。以下，複数工程実験のデータの項目名に関して，簡単な例を示しておく。

原材料をエチレン，プロピレン，ブタンとし，その原料を混ぜて，合成物質を作り，その合成物質を 3 種配合して，配合品を作るという複数工程実験を考える。合成工程では，エチレン濃度，プロピレン濃度，ブタン濃度という 3 つの項目名とする。この段階では，「項目名－項目値」の形になっている。2 次元表形式で書き下すと，表 1 のようになる。

表 1

合成 ID	エチレン濃度	プロピレン濃度	ブタン濃度
合成 1	10	0	50
合成 2	60	30	0
合成 3	20	20	50
合成 4	40	0	50
合成 5	60	30	10
合成 6	10	20	50
：	：	：	：
合成 n	10	20	40

次の配合工程では，合成物質を 3 種配合するので，それぞれの濃度を合成 α 濃度，合成 β 濃度，合成 γ 濃度という 3 つの項目名とすると表 2 のようになる。

表 2

配合 ID	合成 α 濃度	合成 β 濃度	合成 γ 濃度
配合 1	50	0	40
配合 2	60	30	0
配合 3	10	10	70
：	：	：	：
配合 m	10	20	0

この工程だけを見ると「項目名－項目値」の形にはなっている。ただ，このままでは，合成 α 濃度が，合成 1 の濃度なのか？, 合成 2 の濃度なのかはわからないので，困る。そこで，どの合成 ID を配合したのかの情報がわかるように α 合成 ID，β 合成 ID という項目を追加した表 3 のようにする必要がある。

この表の形は，どこかで見覚えがないだろうか？実はこれは，先月の連載記事で，「項目間に論理的関係があり，このままではデータを探すことが困難だったり，X-Y プロットグラフを描くことができない」と説明した形（表 4 と同じ）なのである。

表 3

配合 ID	α 合成 ID	合成 α 濃度	β 合成 ID	合成 β 濃度	γ 合成 ID	合成 γ 濃度
配合 1	合成 1	50	合成 2	0	合成 3	40
配合 2	合成 2	60	合成 3	30	合成 4	0
配合 3	合成 3	10	合成 4	10	合成 5	70
:	:	:	:	:	:	:
配合 m	合成 n	10	合成 1	20	合成 2	0

表 4

実験 ID	原材料名 1	原材料濃度 1	原材料名 2	原材料濃度 2	引張強度
EXP1	エチレン	80	プロピレン	20	10
EXP2	ブタン	75	エチレン	25	12
EXP3	プロピレン	60	ブタン	40	8

先月の連載記事では，「項目間の論理的関係を排除」し，表5のような項目名でデータを記録すべきだと論じている。

今回の複数工程データに関しても，同じように「項目間の論理的関係を排除」すると表 6 のような項目名の表になる。

ある工程の実験を特定する値，つまり「行」に割り振られる ID は，それを参照する下流の工程の項目値として記録するのではなく，項目名の一部に組み込まれないといけないということである。今回の場合は，「合成 1」は，配合工程から参照される合成工程で作成された合成物の ID であり，それをどれだけ使ったかを意味する「濃度」という単語と組み合わされ，配合工程の項目名「合成 1 濃度」になっている。このように参照される工程の ID と連結された項目名（エチレン濃度，合成 1 濃度など）を本連載では，単純な項目名（引張強度など）と区別するために複合項目名と呼ぶことにする。

表 5

実験 ID	エチレン濃度	プロピレン濃度	ブタン濃度	引張強度
EXP1	80	20	0	10
EXP2	25	0	75	12
EXP3	0	60	40	8

表 6

配合 ID	合成 1 濃度	合成 2 濃度	合成 3 濃度	合成 4 濃度	・・・	合成 n 濃度
配合 1	50	0	40	0	・・・	40
配合 2	0	60	30	0	・・・	0
配合 3	0	0	10	10	・・・	70
:	:	:	:	:	:	:
配合 m	20	0	0	0	・・・	10

5 論理的関係のある項目を作ることは，なぜいけないのか？

前章では，「複数工程実験のデータは，どのような項目名で記録すべきか？」と題し，複数工程の実験データの項目名は，どのように決めるべきかを説明した。本章では，「論理的関係のある項目を作ることは，なぜいけないのか？」と題し，論理的関係のある項目名を作ってはいけない理由をデータ分析の観点から論じる。

表 7 の原材料名と原材料濃度は，論理的関係のある項目である。例えば，エチレンの濃度が引張強度にどのような影響を与えるかを調べたいと思ったときに，エチレンの濃度を X 軸，引張強度を Y 軸として，グラフを描きたいと思っても上記表では，非常に面倒だということはわかると思う。また，多変量解析の基礎である重回帰分析をしようと思っても，表 7 のままではいろいろ不都合が生じる。以下，重回帰分析を例にして，どのような点で困るのかを説明していく。

重回帰分析は，原則的には項目値が数値である必要がある。重回帰分析が，y を目的変数，xi を説明変数とした場合に，y = f（x1，x2，x3，x4，・・・），例えば y = ax1 + bx2 + cx3 + dx4 + e（a, b, c, d, e は定数）の形の数式になることを考えると，文字列が入っていると演算ができないので，当然といえば当然である。ただし，世の中のデータ分析をしたいものには，文字列が入るもの（例えば，性別，出身都道府県，業種など）も多い。

表 7

実験 ID	原材料名 1	原材料濃度 1	原材料名 2	原材料濃度 2	合成温度	引張強度
EXP1	エチレン	10	プロピレン	30	100	350
EXP2	プロピレン	30	ブタン	20	200	710
EXP3	ブタン	50	エチレン	10	300	550
EXP4	エチレン	20	プロピレン	40	400	400
EXP5	プロピレン	60	ブタン	20	500	260

実はそういう時のために，ダミー変数化という処理を行うことで，項目値が文字列の項目でも重回帰分析が可能なように拡張できる。ダミー変数化処理とは，表 8 のように項目値が文字列の項目をその項目値を項目名とし，値が 0 もしくは 1 となるように変形することである。

表 8

実験 ID	エチレン 1	プロピレン 1	ブタン 1	原材料濃度 1	エチレン 2	プロピレン 2	ブタン 2	原材料濃度 2	合成温度	引張強度
EXP1	1	0	0	10	0	1	0	30	100	350
EXP2	0	1	0	30	0	0	1	20	200	710
EXP3	0	0	1	50	1	0	0	10	300	550
EXP4	1	0	0	20	0	1	0	40	400	400
EXP5	0	1	0	60	0	0	1	20	500	260

ダミー変数化処理を行うとすべての項目は数値化され，重回帰分析が可能になる。表 8 では，目的変数が引張強度で，説明変数がエチレン 1，プロピレン 1，・・・・温度の 9 変数になる。機械的に行うと表 8 のようになるが，原材料 1，2 が交換可能な対称性を持つという項目間の論理的関係性を考慮すると，表 9 のように変形することも可能である。こうすると説明変数の数は 6 個になる。

表 8 から表 9 への表変形は，（原材料名 1，原材料濃度 1）と（原材料名 2，原材料濃度 2）が交換可能な対称性をもった変数ペアだからできることであり，どのような場合でもこのように変形していいという訳ではない。つまり，データ表を純粋なデータアナリストやデータサイエンティストに渡しただけでは，このような変形さえできないことを理解しておくべきである。

表 9

実験 ID	エチレン	プロピレン	ブタン	原材料濃度 1	原材料濃度 2	合成温度	引張強度
EXP1	1	1	0	10	30	100	350
EXP2	0	1	1	30	20	200	710
EXP3	1	0	1	10	50	300	550
EXP4	1	1	0	20	40	400	400
EXP5	0	1	1	60	20	500	260

交換対称性を考慮した表 9 の 6 個の説明変数でも実は，説明変数としては過剰である。過剰というのはまだ，考慮できていない項目間の論理的関係性があるということである。一般的に，説明変数の数は，その実験の独立変数の数＝実験パラメータの数と一致している必要があり，それよりも多くても，少なくても正しい重回帰分析ができない。正しいデータを分析するためには，項目間の論理的関係性を完全に取り除いた表 10 のような項目にする必要があるのである。

表 10 では，説明変数の数が 4 個になっており，これが真の実験パラメータであり，この実験の独立変数でもある。上表では，変数が 4 個なので，1 次式（y ＝ ax1 ＋ bx2 ＋ cx3 ＋ dx4 ＋ e）を想定すれば，5 個の未定係数を決めることができる。ちなみにこれを解くと，以下のようになる。

引張強度

＝ 10 ×エチレン濃度－ 5 ×プロピレン濃度

＋ 3 ×ブタン濃度＋ 4 ×合成温度

本当の重回帰分析では，誤差も考慮する必要があるので，未定係数の数の十数倍の実験結果が必要であるが，原理的にはこのようにして，目的変数の説明変数依存性を明らかにすることができる。しかし，項目間に論理的関係があると，本当の実験パラメータではない，見せかけの変数が実験パラメータのように扱われてしまい，いくつかの問題により，上記，目的変数の説明変数依存性は導出できなくなってしまう。

表 10

実験 ID	エチレン濃度	プロピレン濃度	ブタン濃度	合成温度	引張強度
EXP1	10	30	0	100	350
EXP2	0	30	20	200	710
EXP3	10	0	50	300	550
EXP4	20	40	0	400	400
EXP5	0	60	20	500	260

一つ目の問題は，項目間に強い相関があると正しく未定係数が決定できないという重回帰分析の性質である。これは，論理的相関でなくとも，結果論として相関が強い場合も問題となるというものである。例えば，年齢，体重，身長などを説明変数とした分析をする場合，体重と身長には，論理的な相関はあり得ない。しかし，実質的に身長の高い人の方が体重は重いという大きな相関があれば，データ分析が正しくできないという問題があり，多重共線性問題と呼ばれている。データ分析における基礎中の基礎の問題である。当然，論理的相関という問題があり，100％の相関であり，重回帰分析において，項目間の論理的相関は絶対に排除しないといけない問題である。

もう一つの問題は，項目間に論理的相関がある説明変数は，独立変数，つまり，真の実験パラメータではないため，どの様な数式を対象数式（モデル式）として採用すべきかが，決めにくくなったり，別途制約条件を設定しなければならないという問題である。

例えば，

引張強度

＝ 10 ×エチレン濃度－ 5 ×プロピレン濃度

＋ 3 ×ブタン濃度＋ 4 ×合成濃度

を，説明変数がエチレン 1，プロピレン 1，・・・・温度の 9 変数になる項目名で表し直してみよう。

引張強度
＝10×（エチレン1×濃度1＋エチレン2×濃度2）
－5×（プロピレン1×濃度1＋プロピレン2×濃度2）＋3×（ブタン1×濃度1＋ブタン2×濃度2）＋4×合成濃度

見てわかる通り，エチレン1×濃度1のように項目間の掛け算が入ってくる。つまり，y＝ax1＋bx2＋cx3＋dx4＋eのような1次式を仮定して，重回帰分析を行うのでは，この解にたどり着かないのである。2次の重回帰分析であれば，エチレン1×濃度1のような項は，入ってくるが未定係数の数は55にも膨れ上がる。その10倍程度なければ，統計的に有意な未定係数が算出できないとすると550実験程度の実験が必要ということになる。また，その時に1次の項である単独のエチレン1や濃度1は，係数は厳密に0になるべきだが，重回帰分析で係数が偶然にも0になることはまずない。明確な形で，「1次の項である単独のエチレン1や濃度1の項の係数は厳密に0になる」という制約条件を課す必要があるのである。

実際には，9変数から6変数にしたような論理的関係をしっかり考え，項目間に論理的関係が全くないような項目にしてから重回帰分析をすれば，見せかけの変数は消失し，正しい分析は可能になる。しかし，一旦，項目を決めてしまうとそれは結構難しい作業になり，データ分析の時にすぐに漏れなく思い出せるものではない。実際，9変数から6変数にするときは，（原材料名1，原材料濃度1）と（原材料名2，原材料濃度2）が交換可能な対称性をもった変数ペアということを考慮することで，見せかけの変数を3つ削減できたが，真の実験パラメータ自由度＝独立変数の数は4であり，さらに2つの変数削減が必要である。しかし，どのような項目間に論理的関係があり，どのように変数削減を行うべきかを明確化するのは非常に難しいのではないだろうか？もし，漏れなく変数削減ができたとしても，データ分析毎に項目の変換が必要となり，データ分析が非常に煩雑になってしまう。そもそも，「データ分析の時にどのような変数削減を行うべきかを明確化できる」ぐらいであれば，最初から項目間に論理的関係がないような項目で実験データ記録をしておくべきであろう。

実は，データを分析し易い項目名とは，実験の独立変数を意識した項目のことであり，実験の独立変数に根差した項目はおのずとデータ分析がし易くなる。当然のことだが，項目間に論理的関係があるということは，その項目は独立変数にはなりえないので，何が独立変数なのかということを意識してデータ記録をすればいいわけである。実際，慣れてしまえば，実験の独立変数に根差した項目はデータ入力としても，違和感はないはずなので，データ共有を目指すのであれば，実験の独立変数とは何かをもう一度見つめ直して欲しい。

6 記録するデータの項目名には，分析で使うべき単位をつけるべき

前章では，「論理的関係のある項目を作ることは，なぜいけないのか？」に関して，説明をした。本章では，「記録するデータの項目名には，分析で使うべき単位をつけるべき」と題し，記録するデータの項目につける単位はどうあるべきかについて，論じる。

例えば，エチレン，プロピレン，ブタン，触媒を混ぜて合成し，その合成物の粘度を測定するだけの単純な工程を考える。項目間に論理的関係のない一番単純な項目名は，次のようなものになるだろう。

a）エチレン重量（g），プロピレン重量（g），ブタン重量（g），触媒重量（g），粘度（Pa・s）

この項目名で，合成物の粘度の触媒量依存性を調べる場合，どうすればいいか？X軸に触媒重量（g），Y軸に粘度（Pa・s）を設定して，X-Yプロットを描けばいいだけなのでは？と思うかもしれない。実は，データ分析をするためにはそれだけではダメで，X軸以外の実験パラメータ，つまり，エチレン重量（g），プロピレン重量（g），ブタン重量（g）のそれぞれが同じ値のデータごとにデータを分類して，X-Yプロットを別グラフとして描く必要がある。そうしないと，X-Yプロットに触媒重量（g）以外の依存性が紛れ込んでしまい，正しい触媒重量（g）依存性をみることができないのである。

そうすると「エチレン重量（g）＝プロピレン重量（g）＝ブタン重量（g）＝100g」の実験と「エチレン重量（g）＝プロピレン重量（g）＝ブタン重量（g）＝300g」の実験は，別グラフのプロットになってしまう。しかし，すべてエチレン，プロピレン，ブタンの比率は同じなので，量効果（容器壁面，表面効果など）が無視できると仮定するなら，粘度など生成物の物理特性は同じになるはずであり，同じグラフに実験結果をプロットしたいはずである。もちろん，エチレン，プロピレン，ブタン

の比率が同じものを何らかの方法で分類をして，それを1つのグラフにプロットすればいいのだが，データ数や項目数が多くなると非常に手間のかかる作業であり，その作業内で間違いを起こしてしまうとデータ分析が台無しになってしまう。実際に，実験データの記録は間違っていなくとも，その後のデータ処理時の変数値の四則演算などで間違いを起こし，データ分析を何度もやり直した経験のある方も多いのではないだろうか？現場の研究者はあまり意識していないことが多いが，データ分析時に煩雑な手動処理を避けることは，データ分析の再現性を確保し，データ分析の信頼性を高めるためには実は非常に重要なことなのである。

容器壁面，表面効果などの量効果はほとんど効かないことが多く，量効果を分離した分析をすることが多い場合は，総重量を分母にした重量百分率の単位の項目が便利であろう。

b）エチレン濃度（wt%），プロピレン濃度（wt%），ブタン濃度（wt%），触媒濃度（wt%），総重量（g），粘度（Pa・s）

このような単位の項目にすると，X軸に総重量（g）を設定し，エチレン濃度（wt%），プロピレン濃度（wt%），ブタン濃度（wt%），触媒濃度（wt%）がそれぞれ一定値のデータで，グラフを描くと総重量依存性が簡単に確認でき，便利である。そして，エチレン濃度（wt%），プロピレン濃度（wt%），ブタン濃度（wt%）のそれぞれが同じ値のデータごとにグラフを描くとすると「エチレン重量（g）＝プロピレン重量（g）＝ブタン重量（g）＝ 100g」と「エチレン重量（g）＝プロピレン重量（g）＝ブタン重量（g）＝ 300g」は，同一グラフにプロットすることができ，a）での問題は解消される。

しかし，よく考えるとエチレン濃度（wt%）＋プロピレン濃度（wt%）＋ブタン濃度（wt%）＋触媒濃度（wt%）＝ 100 であるので，エチレン濃度（wt%），プロピレン濃度（wt%），ブタン濃度（wt%）のそれぞれが同じ値（例えば，それぞれ 15,40,42 wt%）だとすると，触媒濃度（wt%）の値は1つの値（3 wt%）しかとらなくなってしまい，X軸に触媒濃度を設定し，触媒濃度（wt%）依存性を確認しようとしても，プロットは縦に並ぶだけ（触媒濃度＝X値は，1つの値だけ）になってしまう。ちなみに，重回帰分析を行う場合は，濃度のどれか一つ（例えば，エチレン濃度）を変数から外して，つまり，一つの変数を従属変数とし，独立変数の数に合わせて分析すれば，モデル式が適切かどうかということ以外は，問題は発生しない。あくまで，X-Y プロットを描いて，データ分析をしようとするときに問題が出るということである。

次に「触媒濃度（wt%）依存性を確認しようとしても，プロットは縦に並ぶだけになってしまう」問題の解消を試みてみよう。エチレン＋プロピレン＋ブタンの重量を分母にした重量百分率を’wt%と書くこととすると，以下のような項目名が定義できる。

c）エチレン濃度（‘wt%），プロピレン濃度（‘wt%），ブタン濃度（‘wt%），触媒濃度（‘wt%），総重量（g），粘度（Pa・s）

このような単位の項目にすると，エチレン濃度（‘wt%）＋プロピレン濃度（‘wt%）＋ブタン濃度（‘wt%）＝ 100 であり，触媒濃度（‘wt%）はそれらと完全に独立になるので，X軸に触媒濃度を設定し，触媒濃度（wt%）依存性を確認する場合でも，b）で問題となったようにプロットは縦に並ぶだけになってしまうことはない。もちろん，重回帰分析を行う場合は，エチレン，プロピレン，ブタン濃度のどれか一つ（例えば，エチレン濃度）を変数から外して，分析する必要はある。

本章で伝えたかったことを纏めると，記録蓄積するデータの項目およびその単位は，実験時の記録の取りやすさではなく，実験の独立変数を意識し，さらにデータをどのように分析するのかを具体的に考え，分析に沿った単位でなければならないということである。

参考文献

1）上島豊，他　ケムインフォマティックスにおけるデータ収集の最適化と解析手法，技術情報協会，（2023），p39-74