公益社団法人日本技術士会近畿本部.登録　近畿 PE 技術相談室

主成分分析の解説
その 1　統計計算ソフト「R」の紹介

末浪　憲一（経営工学）

1. はじめに

昨日まで経営を支えていた優秀な商品が、今日明日も優れた商品である保証はどこにもない。各社は、他社よりも優れた商品開発にしのぎを削っている。商品の基礎となる技術は日進月歩で進んでいる。全ての企業は、市場の要求に合致した商品を実現するための努力を払っている。

優れた商品開発に向けては、多くの問題や課題を絞り込み解決する必要があり、「主成分分析」は有効な解決手段の一つである。また、簡単な商品であっても、特性項目、それを構成している部品・材料、加工法、良否を判定する検査法等を考えると、無数の検討項目が出てくる。検討項目を効率的に処理するには、一つの実験で多くの要因を比較検討できる「直交実験」が優れている。

これら 2 つの実施には、統計計算ソフト「R」が有用である。「R」は、高等専門学校や大学の研究実験データの解析にもよく利用されている。「R」を直交実験の解析や、これから普及してくると考えられる主成分分析に用いる技術の習得は、技術者にとって有力な武器になると考えられる。

2. 統計計算ソフト「R」のインストール

「R」は、データ操作、計算およびグラフ表示のための統合されたソフトウエアであり統計処理、基本的な科学計算、グラフ処理などで活用できるソフトである。フリーソフトで、誰もが無料で利用できる。また、高度に拡張可能である。

実行には、メインプログラム「R」とパッケージの「Rcmdr（R コマンダー）」が必要である。

メインプログラム「R」は次の手順でインストールできる。

ブラウザを立ち上げ、http：／／cran．ism．ac．jp を表示させる。

最初の画面の「Download R for windows」をクリック。次画面の「install R for the first time」をクリック、「download R．3.6.1　for windows（32／64bit）」クリック、画面下「実行」をクリック。

「R」のインストールが始まる。途中画面の問いに対してマウスで答える。デスクトップ上にアイコンを作成すること。「完了」でタブを全て閉じる。

次に「Rcmdr」を作成する。デスクトップ上に「R」が出ているので、マウスで右クリック。「管理者として実行」をクリック。R 画面が出てくる。上段のメニュー画面（ファイル、編集･･・ヘルプ」の「パッケージ」をクリック、その下に出てきた「パッケージのインストール」をクリック。パッケージ画面　Secure CRAN mirrors　の「JAPAN（Tokyo）」を選び「OK」。

次にいろいろなソフトが縦に表示されている中から「Rcmdr」を選び「OK」。Rcmdr に含まれるプログラムのインストールが始まる。終了すると再度閉じて、「R」の画面を呼び出す。

1ibrary（Rcmdr）と入力してエンターキーを押す。

分析する元データは「Excel」上で分類管理する。必要に応じて「R」にインポートして解析すれば良い。

3, 主成分分析

「主成分分析」とは、統計学上のデータ解析手法の一つで、多くの量的な説明変数を、より少ない指標や合成変数に要約する手法であり、この要約は「次元の縮約」という表現で呼ばれることもある。要約した合成変数のことを「主成分」と呼ぶ。

例えばアンケート調査において、仮に調査項目数が 3 つ以下ならグラフ等で読み取れるが、そ

れ以上の項目数では、特に項目間に相関関係があると、そのままでは理解しづらい。しかし、主成分分析を行うことにより、データの持つ情報をできる限り損なわず、かつデータ全体の雰囲気を可視化し、誰もが理解しやすい形にすることが可能である。

主成分分析は、マーケテイングや研究開発など様々な分野で使われてきており、例えば下記のような活用方法がある。

（1）アンケート調査の結果分析で活用

主成分分析で最も多い活用方法は、顧客満足度調査やブランドイメージ調査、利用者調査などのアンケート結果の分析である。

数多くの質問に対する顧客の回答データを主成分分析し、総合評価を出したり、顧客が重視している点を推測したりするといった使い方である。さらに、主成分得点を基に顧客セグメントがどういう趣向なのかというポートフォリオを作成したり、第 2 主成分以降の内容から顧客が評価している軸を探って商品開発に活かしたりといった活用も進んでいる。また、被験者の評価を数値を用いて行うことができるので、適正や進路などの育成面でも活用できる。

（2）メディアの企業や商品評価で活用

新聞やテレビなどのメディアで掲観される企業ランキングは、一般的には主成分分析での総合指標によって評価されている。例えば、某経済新聞の環境経営度調査や品質経営度調査でも、主成分分析の主成分によるランキングが採用されている。

（3）研究開発で活用

研究開発は数多くの材料を使って実験を行い、膨大なデータが蓄積されていく。それを分析する際にも主成分分析が使われる。例えばお酒造りの際、様々な酵母を様々な条件で試した実験結果を主成分分析すると、総合点の他に、酵母のどういう特性が存在するか、また他の多変量分析と併せてどの特性がお酒の味に好影響を及ぼすのかを推測できるようになる。

4. 直交実験計画法

商品のある特性を改良する場合でも、商品を構成している部品材料、加工条件など多くの検討項目（要因）が考えられる。同じ条件でこれらの要因の比較が必要になる。

取り上げた要因の全ての水準の組み合わせについて実験する要因実験では、因子の数が増えると急激に実験回数が増える。しかし、技術的に解釈の難しい高次の交互作用や、固有技術から存在しないとわかっている交互作用を無視すると、全ての水準の組み合わせの実験をする必要がなく、一部分だけの実験（一部実施法）で済ませることができる。これには直交実験が有効である。

直交表による実験では、2 水準系と 3 水準系の実験がある。

近畿 PE 技術相談室 HP 内の「工程改善実験法今昔」で、2 水準系実験と 3 水準系実験について詳細な実施例を説明している。

https：//kinkipesodan．xsrv．jp／gijyutu－kaisetu／koutei－kaizen－jikkenhou.pdf

2 水準系：マイクロスイッチの温度上昇防止法の実験。

3 水準系：ある電気製品の特性改善の研究

2 つの事例では、元のデータを Excel で管理し、統計ソフト「R・Rcmdr」にインポートして解析している。筆算による従来の解析法と「R」による解析法を比較している。「R」では計算間違いもなく、結果のグラフ化も全て可能であった。

Excel で管理している元データを、定まった方法で「R・Rcmdr」にインポートする。この方法は主成分解析でも利用できた。

その２　主成分分析の実施例

５　東京都内 20 ホテルについて行った主成分分析例

東京都内のホテル２０施設を選び、６つの項目（客室、施設、食事、フロント、サービス、予約対応）について、評価した結果が次のデータ表である。各項目８０点満点で、宿泊客のアンケート結果をもとに点数付けしている。このデータを主成分分析により解析する。

（主成分分析の基本と活用　内田 治著 日科技連 P18 図 2.1 ホテルに関するデータ）

ホテル番号	客室	施設	食事	フロント	サービス	予約対応
1	36	55	41	44	31	53
2	41	65	53	51	44	72
3	20	46	47	56	32	48
4	32	56	50	64	41	54
5	34	54	58	53	29	41
6	34	61	50	50	38	56
7	14	44	30	71	49	72
8	30	52	50	72	51	66
9	42	61	68	66	48	63
10	41	57	56	65	49	65
11	57	67	70	57	42	59
12	14	39	36	42	28	36
13	27	48	84	54	56	56
14	18	42	42	57	47	54
15	41	79	68	62	51	68
16	19	51	39	54	59	64
17	26	59	50	56	31	47
18	25	57	48	62	45	66
19	54	73	67	57	44	67
20	40	61	60	71	58	66

手順１　本データ表を Excel で作成した後、着色部分をクリックボード化（コピー）

手順２　統計ソフト「R. Rcmdr」インストール

手順３　「Rcmdr」に上記ホテルのデータをインストール

「Rcmdr のメニューバー」の「データ」「データのインポート」「テキストファイルまたはクリックボードから」。「クリックボード」「YES」にチェック「OK」、「Rcmdr」メニューバーの「データセット」をクリック。データが正しくインストールされていることを確認

手順 4　主成分分析　６つの項目のデータは、その最大値と最小値に差がある場合、それぞれ基準化した値を用いて、分析する。

基準値＝[（測定データ）－(平均値）]/（測定データの標準偏差）

「Rcmdr」のメニュー蘭で、「統計量」「次元解析」「主成分分析」でウインドウを開き、変数枠で全ての変数「六項目」を選択、オプションのタブをクリックして「相関行列の分析」「スクリーンプロット」「データセット」に主成分得点を保存」にチェックを入れて。「ＯＫ」をクリック。

直ちに下記の解析結果が得られる。（注：今回の分析では、「相関行列の分析」にチェックを入れているから、6 つのデータをそれぞれ PC 内で基準化した後、主成分分析している。）

手順５　分析結果

主成分分析では、固有値（＝分散）が重要である。２０ホテルの六項目の測定値を基準値に変換して主成分分析したから、分散の合計値は６。第１主成分（Comp1)から第６主成分（Comp6)）まで、分散は徐々に小さくなっている。

Component variances：固有値:（＝分散）

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
固有値(＝分散)	3.02088631	1.73143669	0.59716704	0.43150395	0.13852169	0.08048433

各主成分の寄与率(固有値／6)　と累積寄与率

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Proportion of Variance	0.5034811	0.2885728	0.09952784	0.07191732	0.02308695	0.01341406
寄与率（PV/6　%）	50.35%	28.85%	9.95%	7.19%	2.31%	1.34%
累積寄与率	50.35%	79.20%	89.16%	96.34%	98.66%	100%
標準偏差 $\sqrt{CV}$	1.7380697	1.3158407	0.77276584	0.65688960	0.37218502	0.28369761

場合によって異なるが、累積寄与率が 80～90%程度で十分（この場合は第２主成分まで）と考えられている。（１００%に満たない分散＝100－79.2%）が、情報損失の割合）

各主成分における測定項目の固有ベクトル：Component loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
サービス	0.383585	0.4613035	0.33648924	0.468497149	0.00320491	0.55437929
フロント	0.331944	0.4676799	0.05129812	-0.814462168	-0.06011334	0.03866159
客 室	0.430951	-0.4439085	-0.16008879	-0.128605651	0.70749735	0.27295714
施設	0.449772	-0.3620791	-0.39754443	0.011001427	-0.67629044	0.22599135
食事	0.387781	-0.3469856	0.73302717	-0.005525323	-0.12736394	-0.41910070
予約対応	0.451979	0.3462940	-0.40391170	0.316971331	0.14910831	-0.62445640

「Rcmdr」のメニュー蘭で、「統計量」「要約」「相関行列」をクリック、相関行列の窓を開く。

PC1,PC2,PC3,PC4,PC5,PC6 を選び OK。次表を得る。

各主成分の固有ベクトル 6 項目の分散は 1 になるが、各主成分間の共分散は 0 である。

	PC1	PC2	PC3	PC4	PC5	PC6
PC1	1.0000e+00	-9.4737e-17	-7.0678e-16	-2.5426e-16	-3.3954e-16	4.6547e-16
PC2	-9.4737e-17	1.0000e+00	3.5525e-16	1.0982e-16	1.4996e-16	-4.2943e-16
PC3	-7.0678e-16	3.5525e-16	1.0000e+00	5.8443e-16	-4.6187e-16	-5.4731e-16
PC4	-2.5426e-16	1.0982e-16	5.8443e-16	1.0000e+00	4.1065e-16	-1.901e-15
PC5	-3.3954e-16	1.4996e-16	-4.6187e-16	4.1065e-16	1.0000e+00	7.0598e-16
PC6	4.6547e-16	-4.2943e-16	-5.4731e-16	-1.9011e-15	7.0598e-16	1.0000e+00

このことから、固有ベクトルは、互いに独立していることが分かる。

手順６　各主成分における主成分スコア得点の計算

ホテル番号１の第１主成分スコア得点の計算

0.431＊客室実測値の基準値＋0.450*施設実測値の基準値＋0.388*食事の実測値の基準値＋0.332*フロント実測値の基準値＋0.384*ｻｰビスの実測値の基準値+0.452*予約対応の実測値の基準値。

他の場合も同じようにして求めることができる。各ホテルについても主成分（Comp1～Com6）ごとの主成分スコア得点を求めることができる。「Rcmdr」のメニュー蘭で「データセットの表示」

をクリックすると全てのスコア得点値を知ることができる（表示値を直接印刷できない）。

手順７　20 ホテルの主成分分析の結果：評価

第１主成分の固有ベクトル値は、フロント 0.33 から予約対応０.45 である。全ての項目の値が大きくなると、第１主成分の値が大きくなることから第１主成分は、「総合的な良さ」を示す変数であると考えられる。

第２主成分の因子負荷量の符号は、フロント・サービス・予約対応の符号が正で、食事・施設・客室の符号が負である。このことから、第２主成分は「人的サービスが良いタイプか、ホテルが提供する物的サービスがよいタイプか」を示す変数であると考えられる。第２主成分の値がおおきいホテルは、人的サービスがよく、第２主成分の値が小さいホテルは、物的サービスが良いという解釈をすることができる。

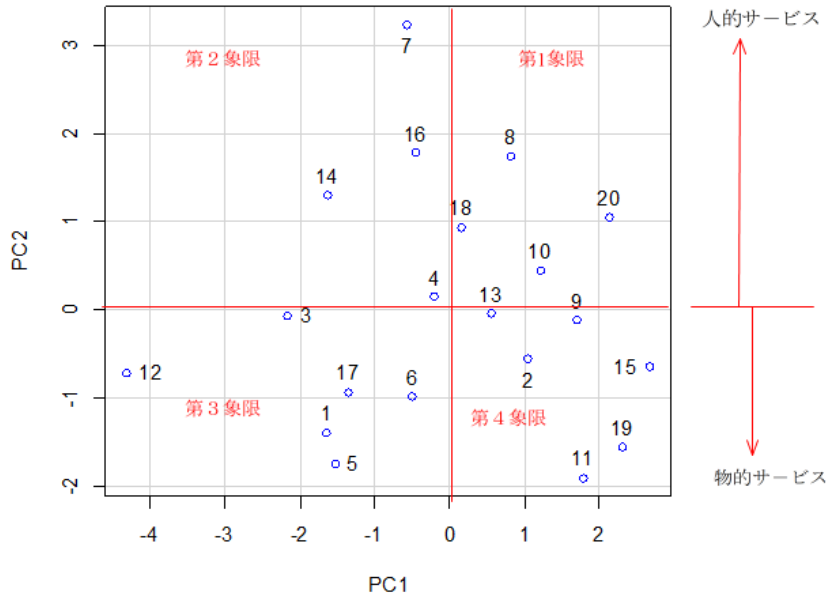
手順８　Comp1 と Comp2 の散布図

Rcmdr のメニューバー」の「グラフ」「散布図」を開き、データ蘭 X 変数に PC1、Y 変数に PC2.

オプションを開き、「マウスインターアクチブ」にチェックを入れる。そして「OK」

次のグラフを得る。各点にマウスを合わせてクリック、打点のホテル番号を得る。

図のように ４象限に区分することで、新しい見方で、各ホテルの比較が可能になる。



手順９　主成分分析で生まれた新しいプロジェクト

２０ホテルについて主成分分析を行った結果、人的（物的）サービスのよいホテルが数値的に明らかになった。どのような点でサービスが良いのか、実地調査を行うことにより、この点が明らかになる。そして、各ホテルにとって新しい目標が生まれて、ホテル業界が活性化されると考えられる。この報告は、例をホテルとして行ったが、新入社員教育に適用すれば、各個人個人に適した教育訓練が可能になると考えられる。（分析の結果は。数値で表現される。）

参考文献　主成分分析の基本と応用。　内田治著。　日科技連出版社

６　おわりに

多変量解析の有効性は認識していたが、数値的な取り扱いが困難であったので近寄りがたかった。最近になって、パソコンが高性能化され、無償の統計計算ソフト「R」を使用できるようになったので、主成分分析などの多変量解析が容易に可能となった。また、大学や高専など学校では、演習データの解析に「R」を使用している。技術者（士）は、そこまでやってきた AI の時代に備えておく心構えが必要であると思う。　　終了

主成分分析解説

主成分分析の基本的な理論は矢野健太郎著　線形代数　第 27 章（固有値と固有ベクトル）で詳細に説明されています。が、基礎知識に十分でない私にとっては、残念ですが、何回復習しても理解できませんでした。そのため、こんなものとして考えることにしました。基本的なことはわからなくても利用していることは沢山あります。例として、マッチはなぜ火が付くのか、など。主成分分析もそのようなものと考えることにしました。

基本は、この式です。

$$AX = \lambda X$$

A はｎ次の正方行列、λは A の固有値、ｎ次の正方行列であるためｎ個の固有値があり、それぞれの固有値には固有値ごとに固有ベクトルが存在します。ｎ個の正方行列ですから、合計ｎ個の固有値に対応した固有ベクトルが有り、互いに独立しています。これだけが、主成分分析の基本であると考えています。

実施例(東京都内 20 ホテル)に適用しますと。ｎは、客室、施設、食事、フロント、サービス、予約対応の 6 個です。従って、6 個の固有値（＝分散の値）と互いに独立な 6 個の固有ベクトルが存在します。そのそれぞれは独立です。固有値の大きさの順番に第 1 主成分から第 6 主成分まで存在することになります。解析の見通しをよくするために、寄与率が小さい主成分固有値（分散＝固有値）は、影響が低いとして、通常は無視します。累積寄与率が 80～90%になる主成分について取り上げます。今回のホテルの例では、累積寄与率が第２主成分で８０%近くになります。

分析の結果、6 個の固有値と固有ベクトルが見つかりましたが、固有ベクトルは何を意味しているのかは、現実の調査の実態に合わせて考えなければなりません。(手順 5)

第 1 主成分は、総合的な評価値です。20 ホテルのスコア得点を求めますと、総合的な評価値で 1 位から 20 位まで厳しいスコア値で順位付けができます。（手順６）

第２主成分以下では、それぞれの６個の項目の固有ベクトルの大きさが与えられています。固有ベクトルの大きさは数学的に求められたもので、実際はどのようなことなのかの判断が必要になります。このホテルの例では、第２主成分の固有ベクトルでは、サービス（0.461）、フロント（0.468）、予約対応（0.346）。客室(-0.443)、施設(-0.362)、食事(-0.347)です。第２主成分では、サービス、フロント、予約対応、の固有ベクトルの係数は（＋）ですが他の項目は（－）ですから、第２主成分の値は、人的サービスのよいホテルの評価値が高くなりますので、人的サービスについての評価であると考えることができます。また、手順６の方法で、主成分ごとに主成分スコア値を求めることができます。

次に、第１主成分と第２主成分の主成分値から散布図（手順８参照）を描くことができます。

第１象限のホテル群は、総合的な評価が高く、かつ　人的サービスも良好なホテル群です。この象限のホテルは、団体客の利用(例えば修学旅行など）に適していると考えられます。

第２象限のホテルは、規模は小さいが、人的サービスが良いホテル群。静かに落ち着いた旅館で、長期滞在して、思考するのに適していると考えられるのではと思います。

第 4 象限のホテル群は、総合的に充実しているが、人的なサービスはそれ程でもない。同じ職場の多くの従業員達の人間関係を良好にするための宿泊・慰安旅行に適していると考えられます。

第 3 象限のホテル群は、総合的な評価が低く際だった特徴も無く、人人から忘れられる存在のホテルであると考えられます。

このようにいろいろなことが考えられますから、原データの作成が重要になります。調査回数を重ねることで、この主成分分析の重要性が認識されるようになると思われます。

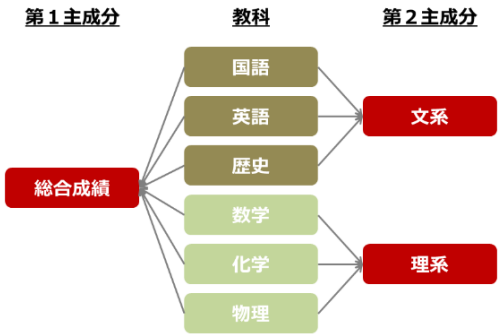
主成分分析とは、次のように説明することもできます。

「主成分分析」とは、統計学上のデータ解析手法のひとつです。たくさんの量的な説明変数を、

より少ない指標や合成変数（複数の変数が合体したもの）に要約する手法。この要約は「次元の縮約」という表現で呼ばれることもあります。要約した合成変数のことを「主成分」と呼びます。
わかりやすく言えば、たくさんの次元（指標）のデータから、全体をわかりやすく見通しの良い 1～3 程度の次元に要約していくことで、たとえば、身長と体重という 2 次元から、BMI（ボディマス指数）という肥満度を表す 1 次元の指標に要約するのが主成分分析、と言えばイメージしやすいと考えられます。

ビッグデータは多変量、多次元であるためそのままでは理解しにくいですが、主成分分析を行うことにより、データの持つ情報をできる限り損なわず、かつデータ全体の雰囲気を可視化し、誰もが理解しやすい形にすることが可能です。

たとえば、６科目のテストを実施している学校があるとして、テスト結果を分析する際、ある教科の点数と別の教科の点数は単純に比較できない。平均点も違えば、点数分布も違うからです。このとき主成分分析を行えば、第 1 主成分に総合成績、第 2 主成分に文系科目／理系科目という指標で、各学生の能力を可視化できます。ある学生の総合的な学力がどのくらいなのか、文系と理系のどちらの能力が高いのかが一目瞭然になるからです。こちらの例を参考にしたモデル図が、下記です。



主成分分析を理解するために必要な用語集

【データの標準化】（手順４）

データの標準化とは、「売上個数」「広告費」「価格」など、様々な単位のデータを扱う際、尺度を揃えて各データの相対的な位置関係を表すために用いる方法で、統計学では通常、平均が 0、分散が 1 となるようにデータを変換することを指します。日常でいうと、偏差値や IQ は、標準化されたデータの一つです。

【固有値】（手順５）

固有値とは各主成分が含んでいる情報の大きさを示す指標です。一般的に「固有値が 1 以上」ある主成分が、元のデータとの関連が深いとされています。

【寄与率】（手順５）

寄与率とは、この主成分だけで元のデータの何割を説明することができているかを表した数字。今回の場合、第 1 主成分で元データの 50%までを説明できていることになります。要約すると必ず漏れてしまう情報があり、そのために第 2 主成分以降が必要となります。

【累積寄与率】（手順５）

第 2、第 3 と続く主成分の各寄与率を足した数値です。　一般的に累積寄与率が 80%以上となるまでの主成分を分析に使います。今回の場合、第 2 主成分までで 80%近くに達したので、ここまでで分析することにしました。

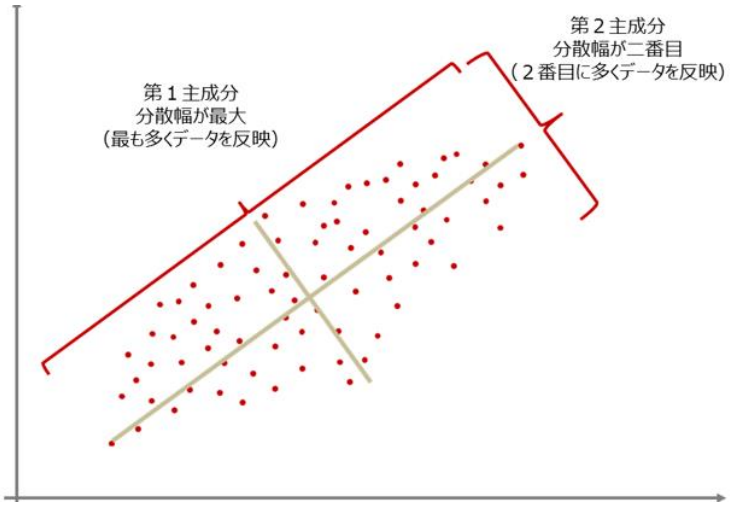
【主成分負荷量】（手順５）

元データの各変数（数値）に対して与えられる係数です。この数値が大きいほど、各変数が主成分に与える影響力が大きいことを表しています。

【主成分得点】(手順６)

元データの各変数を合成した変数となる各主成分を軸とした場合の各変数の得点を表しています。主成分は分析に用いて変数の数だけ出来上がりますが、通常、2 つの主成分を組み合わせた散布図を作成して分析を進めます。各変数の位置関係を見ることで、それぞれの特徴が可視化されます。

▼主成分得点による散布図 （手順６、８）



終わりに

主成分分析の解説は以上ですが、統計解析ソフト「R」と最近のパソコンを用いれば、近寄りがたいと思われた主成分分析は、身近な存在となり、今後益々いろいろな分野で活用されるようになっていくと思っています。そして、この拙文が少しでもお役に立てばと願っています。

以上　2020.03.06

公益社団法人日本技術士会近畿本部登録　近畿 PE 技術相談室
https://kinkipesodan.xsrv.jp/