2022/7/30

コラム名と掲載ページ

掲載ページ	コラム名(先頭の数字は関連する章番号)
(書籍 87pp.)	2.1 特殊なデータ型(NULL, NA, NaN, Inf)
(書籍 88pp.)	2.2 プログラミング時のクイック・レファレンス
2	3.1 離散分布の期待値・平均値と分散
3	3.2 超幾何分布の特徴
4	3.3「事象の独立」の表現
5	3.4 ポアソン分布の特徴
6	4.1 正規性をどうやって確認するか?
7	4.2 統計・確率分野で重要な 2 つの法則:大数の法則と中心極限定理
8	4.3 不偏分散と自由度 n
9	4.4 t 分布とウィリアム・ゴセット
10	4.5 F 分布と分散分析
12	5.1 種々の相関係数と計算式
13	5.2 スケーリング
15	5.3 いろいろな距離
17	5.4 固有値解析と次元圧縮

コラム中にある **R** プログラム**(2** 本**)**

(1)正規性を評価する手法**(6** ページに掲載**)**
［プログラム名 **C_1** ］Rnorm_dist06.R

(2) F 分布の確率密度関数と累積分布関数**(10** ページに掲載**)**
［プログラム名 **C_12**］RFcurve01.R

コラム 3.1　離散分布の期待値・平均値と分散

0 以上 n 以下の整数がそれぞれ出現する確率を $P(k)$とした場合の離散確率分布を考える。これは母集団に対応する(書籍「3.1 事前学習」の節参照)。

この母集団の離散分布につき出現する整数の「期待値」μ は

$$\mu = \sum_{k=0}^{n} k\,P(k)$$

で定義される。

サイコロを振る場合を考えると１から６の目が確率 1/6 で出現するので、「期待値」μ は

$$\mu = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6} = 3.5$$

と計算できる。実験と対応させる場合には、サイコロを振る回数を多くしていけば出た目の「平均値」(標本の平均値)は「期待値」(母集団の平均値)に近づいていく。

また、分布のばらつきの指標となる「分散」を

$$\sigma^2 = \sum_{k=0}^{n} (k-\mu)^2 P(k)$$

と定義する。

サイコロを振る場合を考えると、母集団の分散σ^2は

$$\sigma^2 = (1-3.5)^2\cdot\frac{1}{6} + (2-3.5)^2\cdot\frac{1}{6} + \cdots + (6-3.5)^2\cdot\frac{1}{6} = 2.916\ldots$$

となる。ここで実験と対応させる場合には、サイコロを振る回数を多くしていけば出た目から計算される標本の「分散」は母集団の「分散」に近づいていく。

ここで、わざわざ「母集団」と「標本」を持ち出してきたことには理由がある。まず、全体(「母集団」)と比べて標本数が少ないほど標本から推定される統計量は真(全体)の統計量からはずれる可能性が大きくなることを認識しておかなければならない。次に、標本数が少ない場合に統計量の計算に標本数 n を加味しなければならない。母集団の期待値として標本の平均値(標本数 n)を加味しなければならない。母集団の期待値を標本の平均値(標本数 n で割る操作)で代用することはやむをえないが、母集団の分散 s(不偏分散と呼ぶ)を標本の分散 σ ((標本数 n で割る操作)で代用するのではなく、以下の式の s を使わなければならない。

$$s = \frac{n}{n-1}\sigma$$

n −1 は「自由度」と呼ばれ「不偏推定量」という考え方に基づいている。コラム「不偏分散と自由度 n」も読んでみよう。さらに知りたい場合は統計学の基礎も勉強しよう。

コラム 3.2　超幾何分布の特徴

超幾何分布の確率密度関数、期待値、分散は以下の式から計算することができる。

(1) 超幾何分布の確率密度関数

$$P(k) = \frac{\binom{N_1}{k}\binom{N_0}{n-k}}{\binom{N_1+N_2}{n}}$$

この確率密度関数はフィッシャーの直接確率検定の計算に利用される。

(2) 超幾何分布の期待値

$$\mu = n\frac{N_1}{N_1+N_2}$$

(3) 超幾何分布の分散

$$\sigma^2 = n\frac{N_1N_0(N_1+N_0-n)}{(N_1+N_0)^2(N_1+N_0-1)}$$

コラム 3.3　「事象の独立」の表現

「事象の独立」は高校の数学でも学習する確率の重要な概念であり、以下の式で定義される。

$$P(A \mid B) = P(A \mid \overline{B})$$

となっているとき、A と B は独立である。ここで、左辺は事象 B が起こるという条件のもとで事象 A が起こる条件確率であり、右辺は事象 B でない事象が起こるという条件のもとで事象 A が起こる条件付き確率である。

コラム 3.4　ポアソン分布の特徴

ポアソン分布分布の確率密度関数、期待値、分散は以下の式から計算することができる。

ある期間に平均λ回起こる現象がｋ回起こるとする。

(1) ポアソン分布の確率密度関数

$$P(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

(2) ポアソン分布の期待値

$$\mu = \lambda$$

(3) ポアソン分布の分散

$$\sigma^2 = \lambda$$

ここで、ｅはネイピア数（自然対数の底）。

コラム 4.1　正規性をどうやって確認するか?

正規性の検討方法として、(1)シャピロ・ウィルク検定、(2)Quantile-Quantile(QQ)プロット、(3)コルモゴロフ-スミノフ検定のサンプルプログラムを実行しながら検討してみよう。

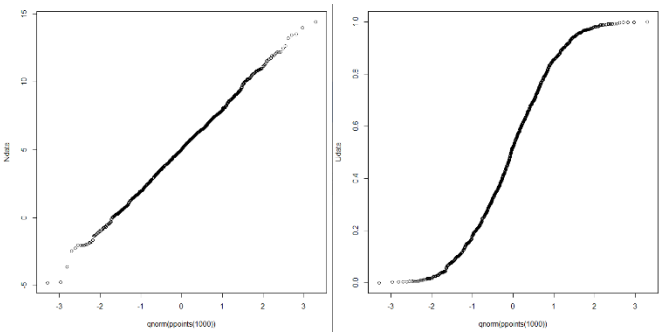
サンプルプログラム　[プログラム C_1]　Rnorm_dist06.R

```
Ndata<-rnorm(1000,mean=5,sd=3)  #Normal random number
Udata<-runif(1000,min=0,max=1)  #Uniform random number
#--Shapiro Wilk test
shapiro.test(Ndata)
shapiro.test(Udata)
#--QQ plot --------------
qqplot(qnorm(ppoints(1000)),Ndata)
qqplot(qnorm(ppoints(1000)),Udata)
#--ks test --------------
ks.test(Ndata,"pnorm", mean=mean(Ndata),sd(Ndata))
ks.test(Udata,"pnorm", mean=mean(Udata),sd(Udata))
```

rnorm 関数は正規乱数、runif 関数は一様乱数を発生する関数であり、それぞれの値はベクトル形式で Ndata、Udata に格納される。これらのデータを shapiro.test 関数に入力すると以下の結果が得られる。

(1)シャピロ・ウィルク検定では、帰無仮説が正規分布であり、対立仮説が正規分布でないと設定されているので、shapiro.test(Ndata)では、p 値が 0.8983 となり正規分布、一方、shapiro.test(Udata)では p 値が $< 2.2e^{-16}$ となり帰無仮説が棄却されて正規分布とみなせない。

(2)qqplot 関数では、直線性が成り立てば正規分布となる。左パネルと右パネルがそれぞれ Ndata, Udata における QQ プロットである。



(3)ks.test は関数、コルモゴロフ-スミノフ検定において正規分布(pnorm)の比較による検定であり、帰無仮説を正規分布とする。Ndata については p 値が 0.9897 、Udata については p 値が 0.001487 となり、Ndata は正規分布、Udata は正規分布とみなせないという結果になる。

今回のテストデータは正規分布(Ndata)と非正規分布(Udata)の間の違いが明確なケースであったが、一般のケースの場合についても(1)～(3)の評価方法を用いて正規性の程度を検討することができる。

コラム 4.2　統計・確率分野で重要な 2 つの法則:大数の法則と中心極限定理

第 4.2 節の学習を終えると正規分布の基本的な特性については十分理解できたと思うが、実際の問題に応用する上でサンプル数が及ぼす影響を知っておくことはとても重要である。統計・確率分野の重要な 2 つの法則についても簡単に学んでおこう。

(1) 大数(たいすう)の法則

「1 つの母集団から、n 個の標本を抜き取りその標本の平均値$\bar{x}$を計算する。もし抜き取り数 n をどんどん大きくして標本の平均値$\bar{x}$をプロットしていくと、その標本平均値は母集団の平均(期待値)に近い値を取っていく。」と表現できる。厳密には n が無限大に漸近する時の振る舞いを弱めの条件(「大数の弱法則」)と強めの条件(「大数の強法則」)に区別することもあるが、この教科書では n が大きい時の平均値の漸近的な振る舞いを利用するだけなのでどちらで理解しても構わない。

知りたい母集団の平均(期待値)は正確には全数検査(連続型データでは無限大になってしまう)して平均を取らなければ求まらないが、十分大きな n であれば標本データの平均と考えてよいことを意味している。母集団が正規分布でなくても当てはまる法則である。実際の応用として、n と平均値の関係のプロットを描くことで母集団の平均値をある精度で推測するために必要な標本数 n の概数を見積もることができる。

(2) 中心極限定理

「知りたい母集団がどのような分布であってもよい。母集団から n 個標本を抜きとり n 個の平均値$\bar{x}$を計算する。復元抽出(第 3.1 節の「事前学習」参照)によって抜き取った後の平均値の計算を何度も繰り返す。繰り返しで得られた平均値$\bar{x}$の集団を分布の形でプロットする。もし母集団が正規分布であればプロットした分布も正規分布となる。もし母集団が正規分布でない場合でも、n の値を大きくしていくことによりプロットした分布は正規分布に近づいていく。」と表現できる。数学的に表現すると、正規分布

$N(\mu, \frac{\sigma^2}{n})$ の形に近づいていくことになる。

正規分布でない分布が平均操作で正規分布に変身していくのは驚きである。人間が暗黙のうちに自然現象のデータあるいは人工的なデータを何らかの形で平均するという操作をしているとすれば、世の中は「正規分布で満ち溢れている」理由が納得できてしまう。

ここで思考実験をしてみよう。もし n が 1 ならばどうなるだろうか。全数検査して一個ずつの値を分布としてプロットすることになるので計算される分布は元の母集団の分布と同じだ。もし n が最初から無限大に近い大きな数だとするとどうなるだろうか。毎回抜き取られるのはほぼ全数検査のデータの集まりなので、1 回ごとの平均値は毎回母集団の平均値と同じになり、その分布は平均値を中心とするデルタ関数のようにシャープに尖った形となる。先に出てきた、正規分布の数式$N(\mu, \frac{\sigma^2}{n})$に n→∞を代入すると、確かにデルタ関数になる。サンプル数 n が違う標本を比較するケースが今後何度か出てくるので、数式に出てくる正規分布の分散が $\frac{\sigma^2}{n}$ の形で表現できることを覚えておいてもらいたい。

コラム 4.3　不偏分散と自由度 n

コラム「離散分布の期待値・平均値と分散」のところで、「母集団の分散」 σ^2 の良い推定値である「標本の分散」の期待値 s^2 (不偏分散と呼ぶ)は各標本の偏差の二乗和をサンプル数 n で割った分散 $\hat{s}^2 = \frac{1}{n}\sum(x-\bar{x})^2$ とはならず、以下の式を使わなければならないことを説明した。

$$s^2 = \frac{n}{n-1}\hat{s}^2$$

カイ二乗分布や t 分布を利用する場合にもサンプル数を n とした場合、自由度は(n-1)の形で表れてくる。これは実は重要な意味を持っている。第 3.1 節の「事前学習」で説明したように、母集団は全数検査でしか全貌が分からない唯一の統計量(期待値)を持つ。「標本」は抜き取る度に異なるサンプルになるので、一般的にはそのたびに異なる統計量が計算される。直観が教えてくれるように、「標本の平均値」を何度も平均した平均値は「母集団の平均値(期待値)」の良い推定量になると考えてよいのでここでは問題にしない。一方、統計の初心者からは必ずと言ってよいほど、サンプル数が n なのになぜ推定や検定で使う分散は分母として n-1 の自由度を使うのかという質問が出る。正規分布のような数式で厳密に表現できる分布を勉強したので、「標本の分散」の平均値と「母集団の分散」の関係を見てから、自由度の意味を考えてみよう。

正規分布の母集団の中から n 個の標本$x_i(i=1,n)$を抜き取ると、この標本の単純な分散は $\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ となる。母集団の期待値(平均値)をμ、標本の平均値を$\bar{x}$として、右辺を次のように展開していく。$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}((x_i-\mu)+(\mu-\bar{x}))^2$　途中で $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$　を使うと、最終的に$\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2-(\mu-\bar{x})^2$ となる。さらに、今展開した$\hat{s}^2$を何度もサンプリングした後の平均値を考える。右辺の第 1 項は「母集団の分散」 σ^2 の定義そのものであり、第 2 項は「標本の平均値」$\bar{x}$の分散であるから、$(\frac{\sigma}{\sqrt{n}})^2$ とも書ける。結局、標本分散$\hat{s}^2$の平均値は　$\hat{s}^2 = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2$ となる。次に、ここで出てくる自由度 n と n-1 の関係を別の視点から考えてみよう。自由度という概念は力学の授業に出てきたのを覚えているだろうか。3 次元空間で m 個の質点が運動する時の運動の自由度は 3 x m で表され、拘束条件を加えるごとに自由度は減っていく。統計分野で分散を考える時も一見自由度はサンプル数 n のように思えるが、分散の計算の中で平均値を使っている。標本 n 個のデータx_iと平均値$\bar{x}$は$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$の関係で結ばれ、1 個の拘束条件が加わっているために自由度が 1 減って、自由度が n から n-1 になっていると解釈できる。サンプル数が大きい場合は n で割っても n-1 で割っても結果はほとんど変わらないことも覚えておこう。

コラム4.4　t分布とウィリアム・ゴセット

ウィリアム・シーリー・ゴセット(William Sealy Gosset)は、ペンネーム「スチューデント (Student)」で論文を書いたため、Student の t 検定として有名になった。ゴセット氏は、オックスフォード大学ニューカレッジで化学と数学を学び、1899 年にギネスビール社のダブリン醸造所に就職した。彼は統計学の知識を醸造と農業（オオムギの改良）の両方に応用しながら実地の研究を重ねた。1906 年から 1907 年にかけてカール・ピアソンの研究室で研究し 1908 年に論文を出した。ギネスでは企業秘密の問題で社員が論文を出すことを禁止していたので、ゴセットは Student というペンネームで論文を発表した。彼のもっとも有名な業績はスチューデントの t 分布と呼ばれる。1908 年の「平均値の誤差の確率分布 (The probable error of a mean)」 をはじめ、ほとんどの論文がピアソンの主宰する Biometrika 誌に発表された。ものづくり企業の技術者が重要な統計分布を発見したことは驚きである。

コラム 5.1　F 分布と分散分析

伝統的な統計の教科書には最後に付録として、必ず「正規分布表」、「カイ二乗分布表」、「t 分布表」、「F 分布表」が掲載されていた。最後の「F 分布表」は F 分布を用いた検定に用いられる。他の手法に比べると利用頻度が多くないことからこの教科書では割愛することになったが、重要な分布の一つになっているので、このコラムで F 分布の概要とその応用分野について簡単に紹介する。

正規分布である母集団からそれぞれ m 個の標本と n の標本を抜き取り、それぞれの分散s_1^2と s_2^2を以下のように計算する。

$$s_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x_1})^2,\ s_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x_2})^2$$

ここで、$\overline{x_1}$と$\overline{x_2}$ は各標本の平均値である。この時、2 つの分散の比は自由度 m と n の F 分布に従うことになる。

実験計画法の分野でよく利用される手法として分散分析がある。例えば 1 つの要因が及ぼす影響を調べたいとする。要因が 2 水準(例えば 2 つの温度で実験データを集める)であれば、第 4.4 節で取り上げた t 分布を利用して 2 つの母集団の差を検定することができるが、3 水準以上になった場合は 2 つずつの検定はできても全部を同時に解析することはできない。一方、これは一元配置分散分析と呼ばれ F 分布を利用した方法で検定することができる。データから計算される F 値が同じ m,n の組に対応する F 分布の有意水準 5%の F 値より大きいか小さいかで仮説検定を行うことになる。

ここでは、F 分布の確率密度関数と累積分布関数を描いてみよう。

例題 C.2　F 分布の確率密度関数と累積分布関数

[プログラム C_2] RFcurve01.R

```
xmin= 0
xmax= 5
par(lwd=2)
curve(pf(x,3,15), xmin,xmax, col="blue",lty=2,main=c("F distribution"),
        xlab="x",ylab="p")
curve(df(x,3,15), xmin,xmax, col="red",lty=1,add=TRUE)
legend(3.5,0.6,c("cumulative","density"),lwd=rep(2,2),col=c(2,4),cex=1.0)
```

[プログラムで利用する関数の説明]

-pf(arg1,arg2,arg3)：F 分布の累積分布関数；arg1=変位値(横軸の範囲)、arg2=自由度 m、arg3=自由度 n。
-df(arg1,arg2,arg3)：F 分布の確率密度関数；arg1=変位値(横軸の範囲)、arg2=自由度 m、arg3=自由度 n。

自由度が m=3 と n=15 の F 分布の場合、プログラムを実行すると確率密度関数(実線)および累積分布関数(破線)は図 C.1 のように表示される。

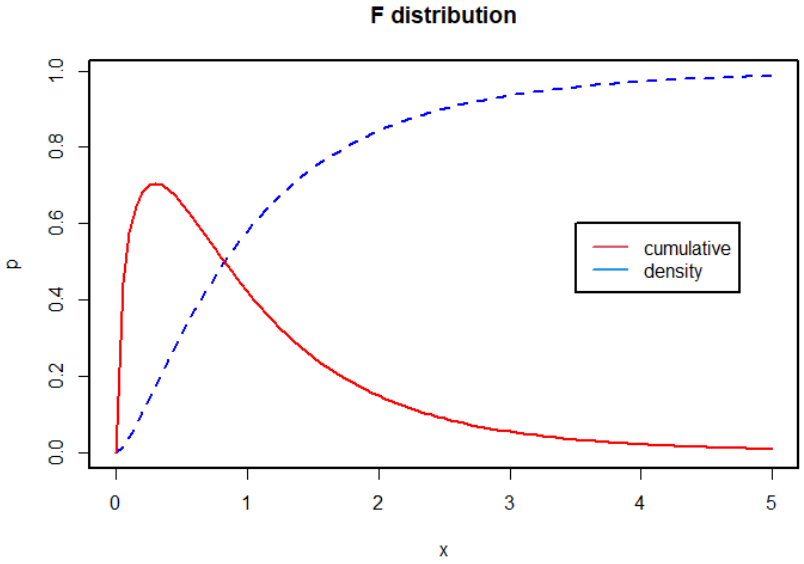


図 C.1　F 分布の確率密度関数と累積分布関数
実線：確率密度関数、破線：累積分布関数

コラム 5.1　種々の相関係数と計算式

名義尺度と順位尺度の場合はピアソンの相関係数を使用することができないため、名義尺度の場合には、φ（ファイ）係数あるいはクラメールの連関係数が用いられ、順位尺度の場合はスピアマンあるいはケンドールの順位相関係数が使用される。

φ（ファイ）係数は 2 変数がともに 2 値データのときに使用され、以下の式で表される。

$$\varphi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{(n_{11} + n_{12})(n_{21} + n_{22})(n_{11} + n_{21})(n_{12} + n_{22}n)}}$$

ここで、xはx_1あるいはx_2のいずれかのみ、yはy_1あるいはy_2のいずれかのみの 2 値しか取らないとし、n_{ij}はx_iとy_jの対のそれぞれの度数を示している。φ係数は第 4 章で学んだ 2×2 クロス表にのみ適用できることに留意してもらいたい。

これに対して、クラメールの連関係数Vは $l \times m$ のクロス表に応用することができ、以下の式で表される。

$$V = \sqrt{\frac{\sum_{i=1}^{l}\sum_{j=1}^{m}(\frac{n_{ij}^2}{n_{i*}n_{*j}} - 1)}{n-1}},$$

$$n = \sum_{i=1}^{l}\sum_{j=1}^{m} n_{ij,}\, n_{i*} = \sum_{k=1}^{m} n_{ik,}\, n_{*j} = \sum_{k=1}^{l} n_{kj,}$$

ここで、変数のi番目のカテゴリーと変数のj番目のカテゴリーに同時に属する度数をn_{ij}としている。カテゴリーの数を 2 とするとφ 係数の場合と同じ結果になる。クラメールの連関係数はφ 係数を一般化したものと考えてよい。

一方、スペアマンあるいはケンドールの順位相関係数は順位尺度を用いる場合に利用される。スペアマンの順位相関係数ρは以下の式で表される。

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n^3 - n}$$

ここで、ペアの数はnであり、それぞれ対応するペアの順位の差はD_iである。ここでは省略するが、同順位がある場合は修正した式が使われる。スペアマンとほぼ同じ傾向を示すと言われているケンドールの順位相関係数τは以下の式で表される。

$$\tau = \frac{K - L}{\binom{n}{2}},$$

ここで、ペアの数はnであり、KあるいはLは項目から 2 項目を選んだときに、順位関係が一致する、あるいは不一致であるペアの数である。

コラム5.2　スケーリング

例題5.3.1.1ではprcomp関数の中でscale=TRUEを指定して、元データにスケーリングという前処理をした。また例題5.2.2.1の中ではscale関数を利用している。スケーリングのメリットはどこにあるのだろうか。

スケーリングには標準化あるいは正規化の2通りあるが、通常は前者を用いる。標準化は平均が0，分散が1になるようにスケーリングし、正規化は最小値を0，最大値を1になるようにスケーリングする。

M個の変数（1,2,…, j,..., j' ,…, M）およびN個の対象（サンプル,1,2,…, i,..., i' ,…N）からなる行列について、多変量解析では通常、各々の変数を平均0分散1にスケーリング（標準化）してから解析を進めることになる。

スケーリング前の行列を、

$$\begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1l} & \cdots & x_{1M} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{il} & \cdots & x_{iM} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i'1} & \cdots & x_{i'j} & \cdots & x_{jl} & \cdots & x_{jM} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{N1} & \cdots & x_{Nj} & \cdots & x_{Nl} & \cdots & x_{NM} \end{pmatrix}$$

スケーリング後の行列を、

$$\begin{pmatrix} x'_{11} & \cdots & x'_{1j} & \cdots & x'_{1j'} & \cdots & x_{'1M} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x'_{i1} & \cdots & x'_{ij} & \cdots & x'_{ij'} & \cdots & x'_{iM} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x'_{i'1} & \cdots & x'_{i'j} & \cdots & x'_{i'j'} & \cdots & x'_{i'M} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x'_{N1} & \cdots & x'_{Nj} & \cdots & x'_{Nj'} & \cdots & x'_{NM} \end{pmatrix}$$

としよう。これらの行列をもとに、スケーリングしたときの共分散とピアソン相関係数を比べると実は同じになる。

いま、列ごとの平均値を、下付き文字 j, j' を使って$\overline{x'_j}$などと表す。また、2つの列ベクトルをx'_jと$x'_{j'}$をベクトル$(x'_{1j}, x'_{2j}, \ldots, x'_{Nj})$と$(x'_{1j'}, x'_{2j'}, \ldots, x'_{Nj'})$で表す。すると、(不偏)共分散cov、ピアソンの相関係数r、(不偏)分散varはそれぞれ以下の式で表される。

$$\mathrm{cov}(x'_j, x'_{j'}) = \frac{\sum_{i=1}^{N}(x'_{ij}-\overline{x'_j})(x'_{ij'}-\overline{x'_{ij'}})}{N-1} = \frac{\sum_{i=1}^{N} x'_{ij}x'_{ij'}}{N-1}$$

$$\mathrm{r}(x'_j, x'_{j'}) = \frac{\sum_{i=1}^{N}(x'_{ij}-\overline{x'_j})(x'_{ij'}-\overline{x'_{ij'}})}{\sqrt{\sum_{i=1}^{N}(x'_{ij}-\overline{x'_j})^2 \sum_{i=1}^{N}(x'_{ij'}-\overline{x'_{j'}})^2}} = \frac{\sum_{i=1}^{N} x'_{ij}x'_{ij'}}{N-1}$$

$$\mathrm{var}(x'_j) = \frac{\sum_{i=1}^{N}\left(x'_{ij'}-\overline{x'_{ij}}\right)^2}{N-1} = \frac{\sum_{i=1}^{N} x'^{2}_{ij}}{N-1} = 1$$

スケーリングをした時の変数間のユークリッドの二乗距離は

$$d(x'_j, x'_{j'}) = \sum_{i=1}^{N} (x'_j - x'_{j'})^2 = \sum_{i=1}^{N} {x'_{ij}}^2 + \sum_{i=1}^{N} {x'_{ij'}}^2 - 2\sum_{i=1}^{N} x'_{ij} x'_{ij'}$$

$$= (\mathrm{N}-1) + (\mathrm{N}-1) - 2(\mathrm{N}-1)\mathrm{r}(x'_j, x'_{j'}) = 2(\mathrm{N}-1)(1 - \mathrm{r}(x'_j, x'_{j'}))$$

となり、実は、(1-相関係数) に比例する。スケーリングすることにより、変数間のピアソン相関と距離を直接関係づけることができる。

平均０分散１にスケーリング(autoscaling)すると、

(1) それぞれの変数のダイナミックレンジを分散として等しくなる、

(2) ピアソン相関係数とユークリッド距離の二乗が負の線形関係になる。

さらに、「対数変換」という前処理もある。もとの変数の値の対数を取ることにより変換する。変数のダイナミックレンジが非常に広く、対数をとることによりデータの分布の正規性が改善されるときには、これを使うといい。

コラム 5.3　いろいろな距離

第 5.1 節の「事前学習」では、データの関係を調べるものさしとして「類似度」と「距離(親近度)」の 2 通りあることを説明し、前者の「類似度」の考え方から第 5.2 節の「相関分析」へとストーリーを展開した。一方、「距離」の考え方は第 5.3.2 項の「多次元尺度構成法」で応用した。「距離」というと高校の数学で学んだユークリッド距離が一般的であるが、データサイエンス分野ではいろいろな距離が利用されている。このコラムでは代表的な距離空間の定義を紹介する。

まず、高校数学で習ったように、3 次元空間上の 2 点 A(a_1, a_2, a_3) と B(b_1, b_2, b_3)のユークリッド距離 d は以下の式で表される。

$$\boldsymbol{d = \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2 + (b_3 - a_3)^2}}$$

データ空間は同一系列のデータ数 N と同じ N 次元空間になるので、N 次元のユークリッド空間上の A($a_{(i)}$; i=1,N) と B($b_{(i)}$; i=1,N)の距離は以下の式に拡張することができる。

$$d = \sqrt{\sum_{i=1}^{N} (b_{\mathrm{i}} - a_{\mathrm{i}})^2}$$

さらに 2 系列間の距離の式を使って、N 系列間の任意の 2 系列間の距離を以下の距離行列 D に拡張することができる。

$$\begin{pmatrix} \boldsymbol{d_{11}} & \cdots & \boldsymbol{d_{1j}} & \cdots & \boldsymbol{d_{1l}} & \cdots & \boldsymbol{d_{1M}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{d_{i1}} & \cdots & \boldsymbol{d_{ij}} & \cdots & \boldsymbol{d_{il}} & \cdots & \boldsymbol{d_{iM}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{d_{i'1}} & \cdots & \boldsymbol{d_{i'j}} & \cdots & \boldsymbol{d_{jl}} & \cdots & \boldsymbol{d_{jM}} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{d_{M1}} & \cdots & \boldsymbol{d_{Mj}} & \cdots & \boldsymbol{d_{Ml}} & \cdots & \boldsymbol{d_{MM}} \end{pmatrix}$$

「(標本)分散共分散行列」C がデータ間の「類似度」関係をベクトルの向きの情報として表現したのに対して、「距離行列」D はデータ間の「距離(親近度)」関係をベクトル間の距離の情報として表現していると言える。「類似度」に出てくるベクトルの向きに関しては $\cos\theta$ の形をそのまま使う以外にいろいろな関数を使うことは可能であるがあまり利用されない。一方、「距離」については通常のユークリッド距離以外の定義もよく利用される。例えば、ユークリッド距離を拡張した「ミンコフスキーのパワー距離」と呼ばれる以下の距離の定義がある。

$$d = \sqrt[1/p]{\sum_{i=1}^{N} (b_{\mathrm{i}} - a_{\mathrm{i}})^{\mathrm{p}}}$$

下図で距離表現の違いを模式的に表しているが、P=2 の場合は通常のユークリッド距離と同じであるが、p=1 と p=∞の場合はそれぞれ市街地距離(あるいはマンハッタン距離)およびドミナンス距離と呼ばれ、データ解析で利用されることがある。「距離」の定義を変えると、データ分析の結果が変わることがある。物理現象を扱っているわけではないので、分析対象のデータ間の関係をうまく表現できる距離を任意に選んで良いが、結果を解釈する上ではデータの「親近度」をどう考えるかのイメージを持って距離の定義を選択することが望ましい。

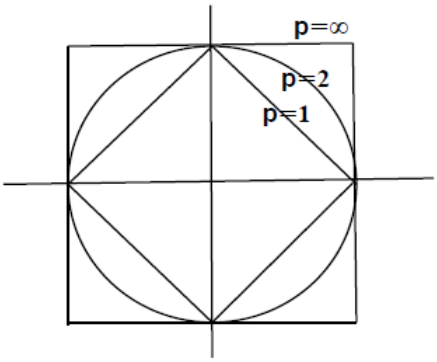


図 C.2　ミンコフスキーのパワー距離

コラム 5.4　固有値解析と次元圧縮

第 5.1 節の「事前学習」およびコラム「いろいろな距離」を学ぶと、データ間の関係は、「類似度」に着目した「(標本)分散共分散行列」C、あるいはベクトル間の「距離」に着目した「距離行列」という数学的な表現となっていることがわかる。しかしながら両方の行列はいずれも高次元(データ系列(変数)の数:M)になっているため、行列の要素を眺めてみてもデータ間の関係を把握することは難しい。次のステップで導入するのは高次元のデータ空間から次元圧縮によって 2 次元あるいは 3 次元のデータ空間に写像し、可視化により理解しやすい形にすることである。特に、ベクトル空間の線形写像(線形変換)を仮定すると、固有値解析を利用できるようになる。

「(標本)分散共分散行列」C の固有値解析が第 5.3.1 項で学ぶ「主成分分析」であり、「距離行列」D の固有値解析が同じく第 5.3.2 項で学ぶ「多次元尺度構成法」である。固有値解析で次元圧縮するとはどういう意味なのだろうか？M 次元空間の系を固有値解析すると、元の M 個の座標軸を回転させて、直交する M 組の固有ベクトルと固有値が答えとして返ってくる。次元圧縮するためには、座標軸をどう回転させるかが重要なポイントになる。

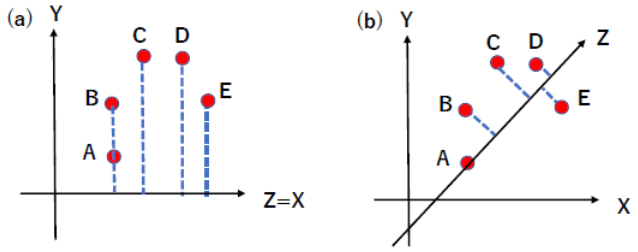


図 C.3　新しい軸とデータとの位置関係

上図のように、2 次元空間上にデータが 5 点あり、次元圧縮して 1 次元にする例を考えよう。1 次元の新しい座標軸を Z とすると、Z 軸の目盛りだけですべてのデータが表現(識別)できることが望ましい。左図(a)のように元の X 軸を Z 軸として Y 軸を使わないとすると、それぞれのデータから Z 軸に垂線の足を下ろして Z 軸の目盛りを読み取ることになる。このケースでは A と B を識別できなくなる。一つの望ましい方策はすべてのデータから降ろした垂線の足に対応する目盛りの分散を最大にすることである。(b)がその方策に沿った新しい軸になる。今回の例は 1 つの新しい軸を見つけるだけであったが、さらに 2 つ目の軸を見つけたい場合は 1 つ目の軸に直交するという条件をつけて、2 つ目の軸に下した垂線の足に対応する目盛りの分散を最大になるようにすればよい。その手順を繰り返しながらどんどん新しい軸を追加していくことができる。数学的には、新しい座標軸を古い座標軸の線形和で表現し、束縛条件(線形和を規格化)のもとで分散が最大になる問題として定式化し、ラグランジュの未定乗数法を使って解くと、例えば主成分分析の場合は分散共分散行列の固有値解析をすることと等価になる。M 次元であれば求まる固有ベクトルが新しい座標軸のベクトルを、また固有値が最大化された分散の値となる。値が大きい順番にならべ

た後の固有値をλ_i(i=1,M)としよう。一番大きな固有値(主成分分析では第 1 主成分と呼ぶ)の全体に対する寄与率は、$\lambda_1/(\lambda_i$の合計)となる。寄与率が十分大きければ(例えば 0.8 以上)データの主要な挙動は第 1 主成分だけで説明できることを意味している。第一主成分だけで不十分であれば累積寄与率を計算し所定の基準(例えば 0.8 以上)に達するまで成分数を増やしていく。最終的に打ち止めした成分数が次元圧縮した結果の次元になるというわけである。

「主成分分析」あるいは「多次元尺度構成法」はデータ間の関係を行列の形に変換し、線形写像を利用して次元圧縮する方法であるが、線形写像では次元圧縮しにくい場合はどのようにしたら良いだろうか。1 つの手段は「自己組織化」である。詳細は第 5.3.3 項「自己組織化マップ(SOM)」を学習してもらいたい。