SAS Technical News

Spring 2006

CONTENTS

1 SAS®によるデータ解析の第一歩
9 Q&A
14 新刊マニュアルのお知らせ
15 SASトレーニングのお知らせ
16 最新リリース情報
16 SAS Technical News送付についてのご案内

特集

SAS®によるデータ解析の第一歩

1. はじめに

データは蓄積しているのだけれども、そこからどのようにしたら有益な情報を得られるかという切実な問題はいつも発生するものです。また、事前に綿密な計画をたててデータを集め分析を行なったときにも、その過程では通常様々な問題が発生します。およそ30年前に統計解析のソフトウェアとして誕生したSASには、高度な分析を行なう機能が多数備わっていますが、分析の前にデータの様子を捉えることは非常に重要です。今号の特集では、Base SAS®やSAS/STAT®ソフトウェアに含まれる分析機能を利用して、「数値データの姿」を確認するいくつかの方法をご紹介します。

2. 平均・再考

はじめに、ある店の顧客が商品を買うために使うお金のことを考えてみましょう。同一の顧客でも、購入金額はそのつど異なるものです。また、顧客間にも購買金額の差異はあることでしょう。

次に実験から得られたデータを思い浮かべてください。こうなるはずだ、という値が仮に存在するとしても、通常はその通りの数値は得られないことでしょう。また、時間を追って観測したデータは、何らかの変化があるはずです。もしある項目の値全てが同じであったときには、本質的にはデータ分析から離れた問題、たとえば計測器の故障で毎回同じ数値が得られていたのかもしれません。男性にのみアンケートを行なった集計結果に男性を表すフラグを追加しても、分析の観点からするとそのフラグはほとんど意味のないものになります。

やや専門的な言葉では、データがばらついた状態のことを「分布」と呼びます。

分布の例

・A高校における学力試験の点数
・ある会社における従業員の年齢
・20代男性における1月あたりの携帯電話の使用料金

また、データを分析しているときには、特徴的な形をした分布が頻繁に現れます。それらには、特定の名前が与えられています。

分布名の例

・ 正規分布（normal distribution）
・ 2項分布（binomial distribution）
・ ポアソン分布（Poisson distribution）
・ ワイブル分布（Weibull distribution）

なお、正規分布はガウス分布（Gaussian distribution）と呼ばれるときもあります。また、ポアソン、ワイブル、およびガウスはいずれも人名です。

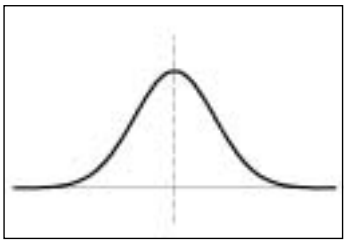


図1 正規分布の形。真ん中あたりにデータが多く、また左右対称

ばらつきがある観測されたデータに直面した場合、そのデータを１つの数値を用いて特徴的に表現することがあります。たとえば、平均という言葉は日常においてよく目にしたり耳にしたりします。

平均の例
- A高校における学力試験の点数の平均
- ある会社における従業員の年齢の平均
- 20代男性における1月あたりの携帯電話使用料金の平均

また、平均を用いて何らかの判断や意思決定を行なうこともよくあります。

続・平均の例
- A君が23分、B君が19分、C君が18分でやり終えたそうだ。平均すると20分だから、それくらいの時間を見積もっておこう。
- 男性の平均寿命は78.64年、女性は85.59年と発表された（厚生労働省発表の平成16年簡易生命表より）。女性の方がおよそ7年長生きということだ。
- 1世帯あたりの平均貯蓄額は約1600万円と新聞で報道されていました。うちはそれよりかなり少ないか。

このように、現実社会では平均は頻繁に使用されているものであり、また極めて重要な統計的指標です。
平均のように、分布をもつデータを特徴的に表す数値は、分布を１つの数値で代表するという意味から、統計の言葉では代表値と呼ばれます。データの「中心」を表す意味での代表値として、平均以外にも中央値や最頻値などがあります。中央値は、データを大きさの順に並べてちょうど真ん中の値であり、中位値やメディアン（median）と呼ばれることもあります。一方、最頻値は観測数が最も多い値のことを指し、モード（mode）とも呼ばれます。
代表値が平均だけであれば非常に楽ですが、なぜ複数の代表値が存在するのでしょうか?その問題に対しては、後の第3項で触れることにして、ここではもう少し平均について考えてみましょう。
自分を含めた10人に対して、満点が100点であるテストを行なったところ、平均が50点、最高点が80点、最低点が20点という結果が得られました。自分の点数が70点であった場合、良い結果であったといえるでしょうか?多くの人にとっては、この点数はかなり良いように感じるかもしれません。
ところが、実際のテスト結果は以下の通りでした。

```
70  80  43  63  20  71  77  31  24  21
↑
自分の点数
```

念のため、Base SASのMEANSプロシジャを使用したプログラムを書いて、平均などを確認してみましょう。このMEANSプロシジャには、データ解析の際に、頻繁に使用される基本的な数値を計算して、出力する機能が備わっています。

プログラム例

```
DATA test;                              /** 分析用のデータを作成 **/
  INPUT score @@;
DATALINES;
70 80 43 63 20 71 77 31 24 21
;
RUN;

PROC MEANS DATA=test;          /** DATA=でデータセット名を指定する **/
  VAR score;                        /** 分析対象の変数を列挙する **/
RUN ;
```

MEANSプロシジャの出力

```
                    MEANS プロシジャ

                    分析変数 : score

 N          平均         標準偏差          最小値          最大値
-----------------------------------------------------------------
10    50.0000000    24.6441339    20.0000000    80.0000000
-----------------------------------------------------------------
```

既知の情報と同じ結果が、平均、最小値、最大値に対して得られていることがわかります。また、後で再度触れますがばらつきに関する指標である標準偏差も併せて出力されています。次にBase SASのRANKプロシジャを用いて自分の点数の順位を調べてみましょう。

RANKプロシジャのプログラム

```
                                /*データセット rankout に結果を出力*/
PROC RANK DATA=test OUT=rankout DESCENDING;
  VAR score;
  RANKS score_rank;             /*順位は変数 score_rank に入る*/
RUN;
PROC PRINT DATA=rankout;          /*作成されたデータセットをプリント*/
RUN;
```

順位の出力

```
                 score_
OBS     score     rank
  1       70        4
  2       80        1
  3       43        6
  4       63        5
  5       20       10
  6       71        3
  7       77        2
  8       31        7
  9       24        8
 10       21        9
```

70点という点数は、悪い成績であるとはいえないかもしれませんが、上から数えると4番目であることがわかります。
では、次のようなケースはどうでしょうか? MEANSプロシジャを利用して平均と最小値、最大値を確認してください。また、RANKプロシジャで順位を確認してみましょう。

```
70  80  43  45  51  49  38  42  47  35
↑
自分の点数
```

この結果からすると、70という点数はそれなりに良いものと判断できそうです。前のデータとは、データの散らばり具合、換言すると分布の形状が異なるようです。このように、自分の得点を単に平均と比べるだけでは、誤った判断にいたる可能性があります。
別のケースを考えてみましょう。10人が100点満点のテストを受け、自分の点数が55点であったとします。平均は50点であり、テスト結果が次のようになっていたとします。なお、データは事前に並べ替えてあります。

テストの結果、その1

```
27  29  38  43  47  50  55  64  71  76
                        ↑
                    自分の点数
```

このときには、55点で自分の点数はおおよそ真ん中あたりです。では、次のテスト結果ではどうでしょうか? 平均は同じく50点です。

テストの結果、その2

```
46  47  48  49  50  50  50  51  54  55
                                    ↑
                                自分の点数
```

最も良い点数ですが、全体的にそれほど差がないことから、本当に良かったと言えるでしょうか? どちらにおいても、最大値と最小値の間で各点数が均等に近い状態で存在しているようですが、値をとりうる範囲が全く違います。これらのことからも、「データの散らばりかた」も重要な意味を持ちそうです。データの散らばり具合を表現する数値にも、色々なものがありますが、分散や先に触れた標準偏差が有名です。
このように、適切でない判断・意思決定を誤って行なわないようにするためには、分布の形状を把握し、何らかの方法で様々な角度から調べてみる必要が少なくともありそうです。そのためには、前記の3つの代表値や、ばらつきの指標などが役にたつでしょう。また、グラフを用いて分布を図示して確認することもよく行なわれます。一方、高度な統計的手法を利用するときには、様々な前提条件が一般的に存在し、その条件が満たされているかを確認する必要があります。その場合にも、前述したような点について注意して、分布の様子を調べておかなくてはなりません。
分析を実行する前に簡単な集計を行ない、グラフを描いて分布の様子を把握することは、思ったよりも重要なことです。この試みによって、データ入力の間違いにも気づく場合があります。また、複雑な分析を行なって得られる結果からよりも、簡明かつ直感的な結果を見つけられるかもしれません。

3. 分析のアプローチとデザイン、母集団と標本

具体的な分析を行なう前に、次のような仮想事例を考えてみましょう。それぞれの判断は、正しいものでしょうか?または、どのような点に注意すべきでしょうか?

- 東京に住んでいる人を対象に、商品の購買動向に関する調査・分析を行なったところ、ある有益な結果が導かれました。その結果をもとに、大阪でキャンペーンを行ないました。

- 不特定多数の人が閲覧可能であるアメリカと日本の2つのWEBサイトで、政治に関心があるかというアンケートを行なったところ、それぞれの国で、1万件程度の回答があり、アメリカでは「YES」が70%、日本では「はい」が55%でした。この結果から、日本人よりアメリカ人の方が、政治に対する関心が明らかに高いと結論づけました。

- ある全く新しい検査方法Aが、旧来から使われている方法Bより優れているかを調べようとしました。そのため、十分高度な技術を持つ検査機関を選び、綿密な計画をたてた上で実験を施行し、適切な統計的手法を用いて分析を行ないました。得られた結果からは、新しい検査方法Aの方が優れていることが判明しました。これからは、全国の検査機関で方法Aが行なわれることになります。

- 有名なテレビ番組で、○○○が健康に良いことが紹介されました。実際に○○○を2人が1週間摂取し続ける実験を行なったところ、全く摂取していない他の2人に比べて、ある検査値が平均10%ほど下がったとのことです。この番組で紹介されるとスーパーマーケットから○○○が消えるので、そうなる前に早速明日買っておこうと考えました。

データを分析することによって、何らかの結果を得ようとするとき、ソフトウェアを使用して行なう計算処理の前には、データをどのようにしてどれくらい集めるかを検討したり、また得られたデータを分析できる形に整えるなど、一般に骨の折れる作業が存在します。
たとえば、日本人男性20才の身長の平均を知りたいとします。しかし、全ての人の身長を計測することは現実的に不可能です。そのため、調査法として適切な手法に基づき、20才の日本人男性をある一定数集めて、そこから計算された平均を算出し日本全体における身長の平均を推測・推定するような形になります。統計解析・データ解析の領域では、母集団(population)と標本(sample)という言葉が現れます。母集団とは、分析の対象である集団のことであり、標本はそこから取り出した小さな集団です。前の例では、母集団は20才の日本人男性全てです。一般的には、想定している母集団から相対的に小さな標本を「ランダム」に集め、その標本を用いてあらかじめ決めておいた分析方法を適用して結果を得て、母集団に対しその結果を反映させる、といった図2のような流れになります。

この母集団と標本の区別は重要です。母集団の情報は、通常はほとんど得られていません――母集団の情報が完全にわかっていたとしたら、幸せなことです。なぜなら、全ての情報は今あなたの手元にあるからです――。

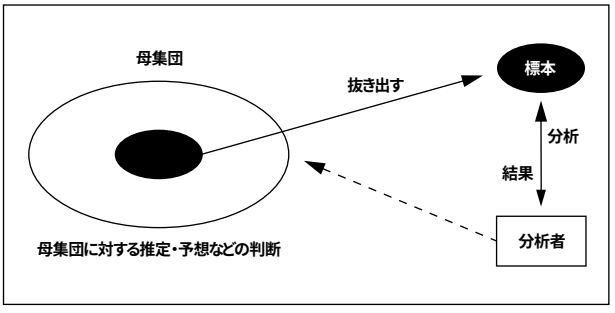


図2 母集団と標本、及び分析に関するイメージ

一方、たとえばデータベースに存在する全ての顧客データを用いて何らかの分析を行ない、その結果に基づいて顧客へアクションを起こすといったケースでも、そのアクションの対象は「過去の顧客」、「顧客の過去」ではなく、「未来の顧客」、「顧客の未来」のはずです。蓄積された既存データに対して後づけで分析を行なう場合には、図3のように母集団をどのようなものと想定できるか、標本はそこからどのように集めた形になるか、また得られた結果をどのあたりまで反映させることができるのか、十分注意する必要があります。図2と図3の違いを考えてみましょう。

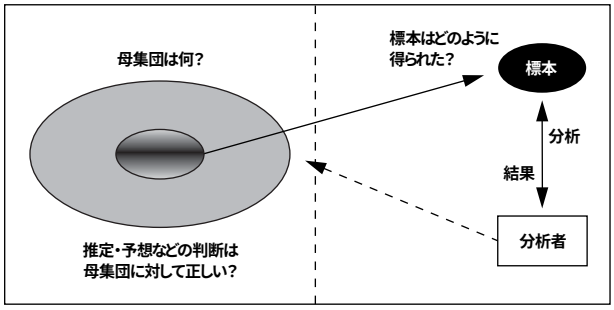


図3 母集団と標本、及び分析に関するイメージ、その2

「堅牢な分析」を行なうためには、これらの点について事前にしっかりとデザインしておく、または意識をしっかりと持っておく必要があるでしょう。もし分析対象の集団（母集団）が判然としておらず、現在持っているデータ（標本）の素性が明らかでなければ、分析を行なった上で仮に良さそうな結果が得られたとしても、どこまで意味のあるものであるかがあやしいものになります。むしろ、誤った判断をしてしまう可能性も高くなるかもしれません。
これらの点に気をつけた上で、本項の冒頭の問いについて再度考えてみてください。また、可能であれば周囲の人と議論してみましょう。

4. データを眺める、要約する、その1

本稿の冒頭で触れた平均貯蓄額について、もう少し調べてみましょう。総務省統計局が公開している「家計調査年報 平成16年 <<貯蓄・負債編>>」によると、日本の2人以上の世帯（母集団に相当します）における貯蓄現在高の平均は1692万円とのことです。（「平成16年の貯蓄・負債の概況」より。）調査によって得られた各貯蓄高別の家庭数は以下の通りです。

貯蓄額	家庭数	貯蓄額	家庭数
～100万円	690	～1200万円	558
～200万円	456	～1400万円	412
～300万円	454	～1600万円	372
～400万円	443	～1800万円	284
～500万円	400	～2000万円	277
～600万円	361	～2500万円	520
～700万円	390	～3000万円	384
～800万円	350	～4000万円	534
～900万円	313	4000万円～	825
～1000万円	295		

表1 貯蓄額と家庭数［出典］ 総務省統計局発表／家計調査年報平成16年 ＜貯蓄・負債編＞ 統計表／第1表 貯蓄・純貯蓄・負債現在高階級別

額が大きくなるほど貯蓄額の幅が広くなっているため、家庭数が最も多いのは4000万円以上のところですが、次に多いのは100万円以下のグループであり、また貯蓄額が増えるにつれて家庭数が減っていく傾向があります。このデータをもとに棒グラフを描いてみます。なお、より正確な判断を行なうために、個々の棒の高さは棒の幅に応じて調整して表示しています。具体的には、幅が200万円のところは、高さを半分、500万円の棒については5分の1にしています。また、公開されているデータからはグラフの一番右側では上限がわからないことから、こちらも意味のあるように調整しています。

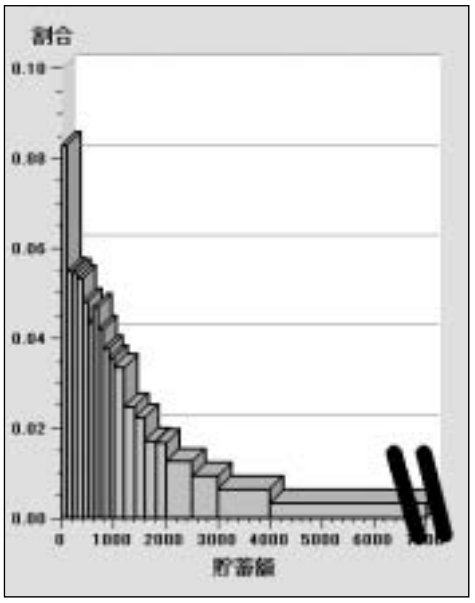


図4 表1のデータをもとに作成した棒グラフ

横方向の軸は貯蓄額を、縦方向の軸は全体に対する割合を表しています。平均である1692万円は、どのあたりになるでしょうか？ 線を引いて確認してみましょう。もし可能であれば、我が家の貯蓄額がどのあたりになるか、また平均と比べて右か左かを考えてみましょう。実は貯蓄現在高に関して中央値もきちんと報告されており、1024万円とのことです。こちらについても線を引いてみましょう。平均と中央値では、どちらが現在貯蓄高の分布を「より正しく」代表していると思いますか?
平均という言葉はあまりにも一般的となっているので、マスメディアが中央値を平均とともに紹介していたとしても、それほど記憶に残らないかもしれません。また、仮に中央値に着目すべきと書かれていても、あまり正確なイメージが沸いてこないかもしれません。このようにグラフを併せて利用することは、一般にわれわれの理解を助けてくれます。

では、最頻値はどのあたりにあるでしょうか？ 図4のグラフから考えてみましょう。平均、中央値とともにグラフに線を入れ、それらの関係を考えてみてください。どの代表値が貯蓄高の分布を「最も正しく」代表しているでしょうか?

5. データを眺める、要約する、その2

別のデータを使用して、SASの機能を用いて簡単な分析を行なってみましょう。次の仮想データは、ある商店のある期間における顧客ごとの売上に関するものです。そのうち、先頭の10オブザベーションだけ表示しています。

サンプルデータ

```
顧客ID 性別 品目数 売り上げの合計(円)
  1      M   2     120000
  2      F   6      70000
  3      F   3      32100
  4      M   1        980
  5      F   5      87150
  6      F   5      46320
  7      F   2      21300
  8      F   2      12000
  9      M   1       3400
 10      F   3      11180
......
```

SASデータセットでは、変数名を順に id, sex, count, totalとしましょう。また、性別ではMが男性を、Fが女性を表しているものとします。前述のMEANSプロシジャを使用しても有益な統計的情報は得られますが、一般に変数に関して詳細な情報を得るためには、UNIVARIATEプロシジャが有用です。UNIVARIATEプロシジャで、数値データである顧客ごとの売り上げの合計について分析してみましょう。

UNIVARIATEプロシジャの指定例

```
PROC UNIVARIATE DATA=sales;
  VAR total;
RUN;
```

UNIVARIATEプロシジャの出力例

UNIVARIATE プロシジャ
変数 : total

モーメント

N	3689	重み変数の合計	3689
平均	42637.5142	合計	157289790
標準偏差	30349.1305	分散	921069723
歪度	1.98872843	尖度	5.61947686
無修正平方和	1.01034E13	修正済平方和	3.39691E12
変動係数	71.1794087	平均の標準誤差	499.679953

基本統計量

位置		ばらつき	
平均	42637.51	標準偏差	30349
中央値	33930.00	分散	921069723
最頻値	26240.00	範囲	232020
		四分位範囲	30830

位置の検定 H0: Mu0=0

検定	--統計量---		-------p 値-------
Student の t 検定	t 85.32965	Pr > |t|	<.0001
符号検定	M 1844.5	Pr >= |M|	<.0001
符号付順位検定	S 3403103	Pr >= |S|	<.0001

分位点 (定義 5)

分位点		推定値
100%	最大値	233000
99%		155430
95%		103260
90%		81430
75%	Q3	53540
50%	中央値	33930
25%	Q1	22710
10%		15000
5%		11570
1%		6770
0%	最小値	980

極値

----最小値----		-----最大値-----	
値	Obs	値	Obs
980	4	221270	1891
3400	9	222610	1179
5070	1695	232710	1973
5080	3341	232840	2394
5210	662	233000	694

UNIVARIATEプロシジャの出力を確認してみましょう。
1つ目のテーブルには、欠損でないオブザベーション数（N）など、データに関する基礎情報に加えて、平均や標準偏差、および分散などが出力されています。
2つ目のテーブルには、3種類の代表値、平均、中央値および最頻値とともに、範囲や四分位範囲などのばらつきの指標が出力されています。
「位置の検定 H0: Mu0=0」テーブルでは、ある仮定のもとでデータの中央値、または平均が0と異なると言えそうか、積極的には言えなさそうかを調べる統計的手法の結果が出力されています。ただし、ここでは必ず0以上である売上高について考えていることから、これらの出力はあまり意味のあるものではないでしょう。
4つ目の「分位点」のテーブルでは、小さい方から数えて全体の何パーセントの位置に存在するか、代表的な数値を選んで表示されたものです。たとえば、75% Q3の行で出力されている数値は、第3四分位点、75パーセント点、またはQ3と呼ばれ、小さい方から数えて75パーセントのところに存在するデータの値です。
また、25% Q1は小さい方から数えて25パーセントのところに存在する点であり、第1四分位点、25パーセント点、またはQ1と呼ばれます。その他のパーセント点も、同じように定義されます。「基本統計量」テーブルで出力されていた四分位範囲は、この第3四分位点と第1四分位点の差です。この数値が範囲（最大値と最小値の差）の中で占める割合が大きなものであれば、データが中心部分に集中して存在しているのではなく、ちらばりが大きいことが示唆されます。

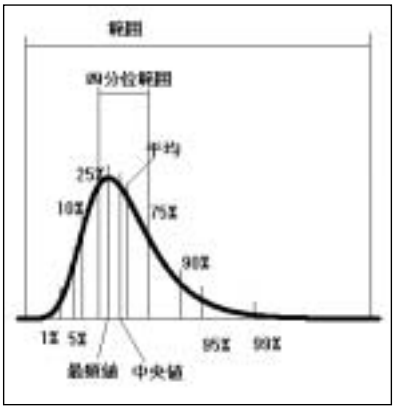


図5 分布における代表値やパーセント点の位置のイメージ

最後の「極値」テーブルでは、最も小さい値が下から5つ、また最も大きな値も上から5つ表示されています。この部分は、裾、またはテールなどと呼ばれ、分布の形状を把握するときには重要な場所の1つとなります。
このデータには性別の情報を含む変数が存在していました。性別が異なれば、消費行動に何らかの違いがあるかもしれません。そこで、性別ごとに変数totalを分析してみましょう。そのためには、UNIVARIATEプロシジャでCLASSステートメントを使用します。

CLASSステートメントの使用例

```
PROC UNIVARIATE DATA=sales;
  CLASS sex;            /*CLASSステートメントで性別の変数sexを指定*/
  VAR total;
RUN;
```

CLASSステートメント指定による出力例（抜粋）

UNIVARIATE プロシジャ
変数 : total
sex = F

基本統計量

位置		ばらつき	
平均	47251.42	標準偏差	32175
中央値	37590.00	分散	1035261909
最頻値	24700.00	範囲	227930
		四分位範囲	33400

分位点 (定義 5)

分位点		推定値
100%	最大値	233000
99%		162630
95%		112300
90%		88100
75%	Q3	59230
50%	中央値	37590
25%	Q1	25830
10%		17860
5%		14820
1%		10400
0%	最小値	5070

極値

----最小値----		-----最大値-----	
値	Obs	値	Obs
5070	1695	221270	1891
5210	662	222610	1179
5220	2288	232710	1973
5450	2495	232840	2394
5910	1753	233000	694

変数 : total
sex = M

基本統計量

位置		ばらつき	
平均	31284.53	標準偏差	21433
中央値	25820.00	分散	459354578
最頻値	14270.00	範囲	197800
		四分位範囲	23240

分位点 (定義 5)

分位点		推定値
100%	最大値	198780
99%		105630
95%		71450
90%		59490
75%	Q3	39720
50%	中央値	25820
25%	Q1	16480
10%		10540
5%		8400
1%		5780
0%	最小値	980

極値

----最小値----		-----最大値-----	
値	Obs	値	Obs
980	4	124200	574
3400	9	127700	3481
5080	3341	146520	2072
5220	1139	150410	1555
5420	794	198780	2429

なお、男性と女性はそれぞれ1066人、2623人でした。性別間では売上高にどのような違いがあるでしょうか? これらの数値にはそれぞれ意味があり、ときとして雄弁に分布の様子を物語ります。しかし、数字ばかりを見ても理解が難しいこともあります。もっとわかりやすくデータの様子を知る方法はないものでしょうか? 次の項では、グラフを描いて調べてみましょう。

6. データをもとにグラフを描いて見よう

一言でグラフを描くといっても、どんなグラフを描けばよいのでしょうか?統計学辞典(東洋経済新報社)の「第III章 汎用的方法」では、10個以上のグラフが紹介されており、実際には目的に応じて使い分ける形になります。ここでは数値データの分布をとらえるという観点から、ヒストグラムと箱ひげ図を描いて分布の様子を確認してみましょう。

ヒストグラム（histogram）とは、観測された数値データの値に応じていくつかの区間に分けて、その区間ごとに対して柱状のグラフを描き、観測されたデータの個数などを表現したものです。UNIVARIATEプロシジャでは、HISTOGRAMステートメントが使用できます。顧客ごとに売れた品目数と売り上げの合計について、それぞれヒストグラムを描いてみましょう。プログラム中のMIDPOINTS=の指定は、各柱の中心点を与えるものです。なお、この指定方法を変えると、ヒストグラムの見栄えが変わることに留意してください。

HISTGRAMステートメントの使用例

```
PROC UNIVARIATE DATA=sales;
  HISTOGRAM count / MIDPOINTS=0 TO 10;
  HISTOGRAM total / MIDPOINTS=0 TO 240000 BY 20000;
RUN;
```

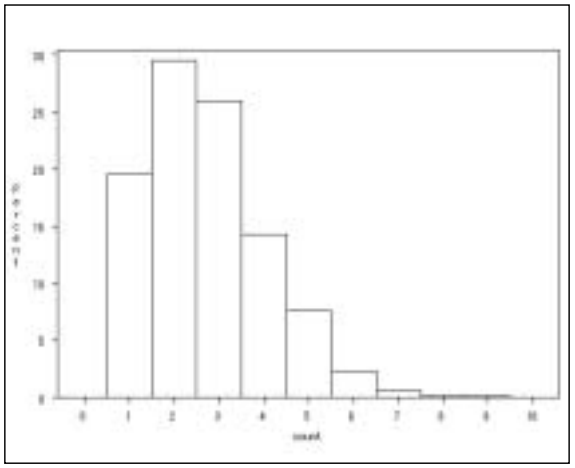


図6 品目数に関するヒストグラムの出力例

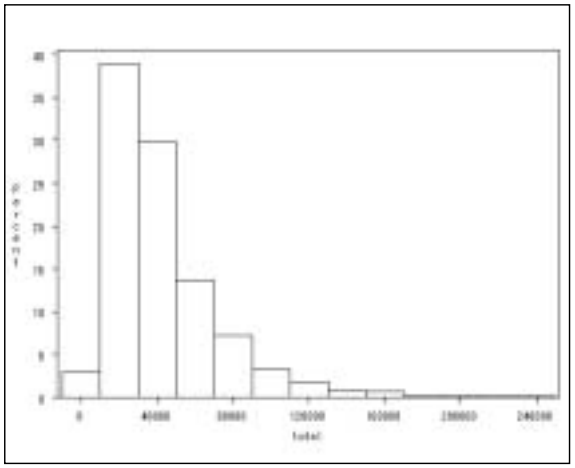


図7 売り上げの合計に関するヒストグラムの出力例

ひとりあたりの購買された品目数は2個のあたりが多く、また売上高については20,000円の柱が最も高いものとなっています。また、いずれも右側に裾が長く伸びた形状となっています。このデータには、性別という分析で利用できそうな情報がありました。そこで、性別ごとにヒストグラムを描いてみましょう。どのようなことがわかるでしょうか?

性別ごとに売上合計のヒストグラムを描く例

```
PROC UNIVARIATE DATA=sales;
  CLASS sex;
  HISTOGRAM total / MIDPOINTS=0 TO 240000 BY 20000;
RUN;
```

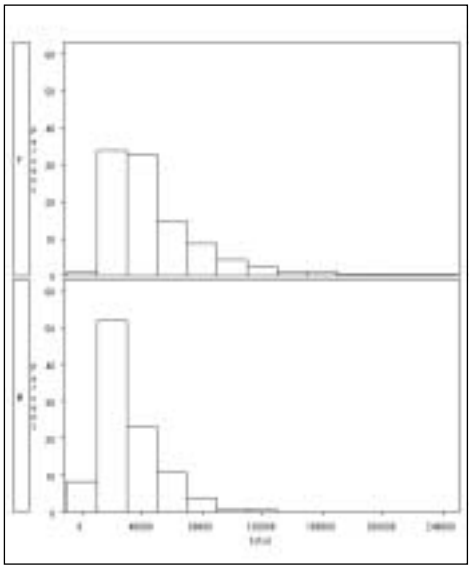


図8 売り上げの合計に関する性別ごとのヒストグラムの例

一方、箱ひげ図(box-and-whisker plots)は、分布の形状を把握するために有用な様々な数値情報を、ある独特な図式表現を用いて表現したものです。ヒストグラムほど直感的ではないため、慣れないとわかりづらいかもしれませんが、ヒストグラム描画時に発生する「棒の幅(前の例では20,000円)をどれくらいに定めるか」といった問題が発生しないというメリットもあります。この箱ひげ図を描くためには、SASでは前述のUNVARIATEプロシジャも利用可能ですが、よりきれいな箱ひげ図を作成するBOXPLOTプロシジャをここでは使用してみましょう。ヒストグラムのときと同様に、男女別で箱ひげ図を描いてみます。

BOXPLOTプロシジャの例1

```
PROC SORT DATA=sales;                    /*性別で事前にソートをしておく*/
  BY sex;
RUN;

PROC BOXPLOT DATA=sales;
  PLOT total*sex;
RUN;
```

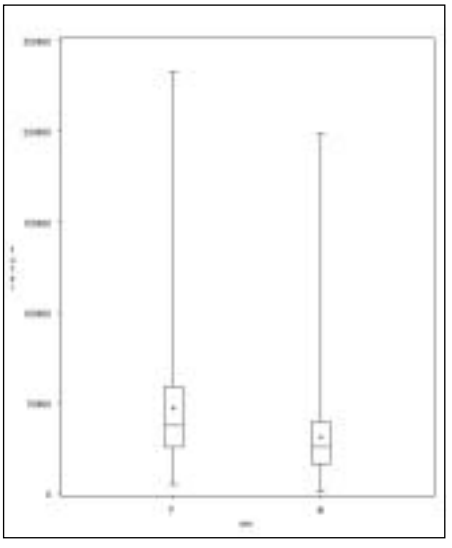


図9 BOXPLOTプロシジャによる出力例、その1

中央に「箱」が描かれ、「ひげ」と呼ばれる線が上は最大値まで、下は最小値まで伸びています。箱の上の線は第3四分位点、小さい方から数えて75%のところに存在するデータの値です。一方、箱の下側の線は第1四分位点（25パーセント点）です。箱の真ん中に横に引かれている直線は中央値を、一方「+」の記号は平均を表しています。第3四分位点から最大値まで伸びるひげは、下側に現れているひげよりかなり長いものであり、これは購入金額の非常に高い顧客が存在していることを示唆しています。このような現象に影響を受けやすい平均は上側に引っ張られていて、中央値よりもいくぶん大きな値となっています。UNIVARIATEプロシジャが出力した平均や分位点の値、およびヒストグラムとこの箱ひげ図を比べて、対応関係を確認してみましょう。
このタイプの箱ひげ図では、第3四分位点から最大値までの間の様子があまりわかりません。そこで、別のタイプの箱ひげ図を描いてみます。

BOXPLOTプロシジャの例2

```
PROC BOXPLOT DATA=sales;
  PLOT total*sex / BOXSTYLE=SCHEMATIC;
RUN;
```

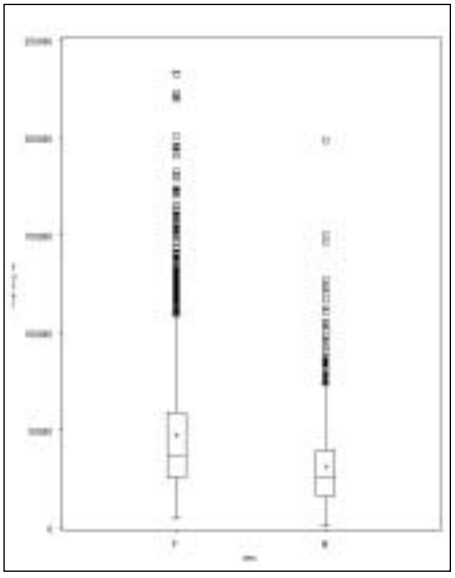


図10 BOXPLOTプロシジャによる出力例、その2

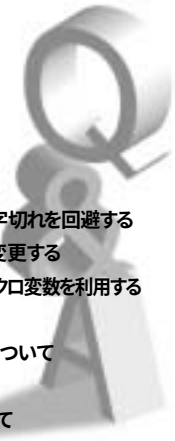
BOXSTYLE=SCHEMATICは、ひげを最大値や最小値まで伸ばすのではなく、一定の値以上離れた点に関しては、各点をプロットして際立たせて表示させる指定です。図10ではプロットが重なっているため、ややわかりづらいですが、この出力では多くの点が上側に表示されており、データの中心部分から大きい方に購入金額の高い顧客が存在していることと、存在の様子を示しています。
データの中心部分から大きく外れた値のことを外れ値（outliers）と呼びます。一般的に、外れ値には色々な意味で注意する必要があります。このケースでは、購買金額の高い顧客が、外れ値として表示されています。これらの優良顧客の購買心を高めることによって、効果的に増収が図れるかもしれません。
また、仮に実験で得られたデータにおいてこのような外れ値があったとすると、なぜそのような値を観測したのか注意深く調べる必要があるでしょう。事前に想定していた分布の形状は正しいものではないのかもしれません。または、結果の入力ミスや、測定機器の故障も考えられます。
SASにはその他にも、SAS/GRAPHに含まれる機能を用いて多種多彩なグラフを描くことができます。また、SAS9.1では評価版ですが「ODS Statistical Graphics」を使用すれば、それぞれの分析において有用なグラフが作成されます。この機能は、次期リリースSAS 9.2において正規版となり、更に大幅に拡張される予定です。

7．おわりに

これまでに紹介したような内容については、どこで知ることができるのでしょうか？ 現在では、WEB上にも大量の情報が公開されており、最もアクセスが簡単です。たとえば、家計の貯蓄データを発表している総務省統計局のWEBページでは、「統計学習サイト」という項目が公開されています。主として小・中・高校生向けのようですが、統計・分析の入り口として役にたつ内容も多く、中央値や最頻値に関する丁寧な解説もあります。また、本稿で引用した貯蓄額に関する調査の際に行なわれた調査方法も、詳しく説明されています。
データ解析における入門者向けの書籍も大いに参考になるでしょう。店頭で手に取って内容を確認し、自分にあった本を探してみてください。なお、英語で書かれたものまで含めると、数多くの書籍があります。最新の理論を扱った書籍は英語のものしか存在しないことも多く、最初から英語で書かれた本に親しむのも良いかもしれません。もちろん、SASに関連した書籍については、弊社でも取り扱っています。また、信頼できる周囲の人に尋ねることも良いでしょう。

これまでに紹介したような内容に基づいただけでも、色々なことが見えてくるともあります。この特集が、お客様の業務の一端になれば幸いです。

Q&A



- REPORTプロシジャで、1行おきに背景色を変える
- IMPORTプロシジャでCSVファイルを読み込む際の文字切れを回避する
- 自動生成されたマクロ変数を利用して変数名を一括変更する
- Access to Oracleのパススルー機能にてWHERE句にマクロ変数を利用する
- データから数値または、文字だけを抽出する方法
- 現在実行しているプログラムのファイル名取得方法について
- 特殊記号を含む文字列を使ったマクロ変数の作成
- AUTOREGプロシジャにおける収束ステータスについて
- 多次元正規分布に従う乱数列を生成する方法について

Q ODS HTMLで、REPORTプロシジャの出力をHTMLファイルに出力しています。出力された行の背景色を1行おきに変えることはできますか。

A **CALL DEFINEステートメントを使用して、特定の行や列のスタイルを指定することができます。この機能とMOD関数の戻り値を組み合わせて、1行おきに背景色を変更できます。**

MOD関数説明

MOD関数は、被除数を除数で割ったときの余りを求めます。

MOD関数の構文

```
MOD(被除数,除数)
```

CALL DEFINEステートメント説明

レポート定義に使用します。STYLE=属性では、CALL DEFINEステートメントの影響を受けるセルに適用するスタイルを指定します。

CALL DEFINEステートメントの構文

```
CALL DEFINE (column-id, 'attribute-name', value);
```

下記の例では、MOD関数にて行数を2で割り奇数と偶数に分類し、1行分のカラムに対するすべての背景色を個別に指定しています。

例:CALL DEFINEステートメントで背景色を指定

```
ODS HTML BODY='c:¥sashtml¥report1.html';
  PROC REPORT DATA=sashelp.class NOWD
  STYLE(HEADER)=[BACKGROUND=CX00ccff];NPUT a1 a2 b1 b2 x $ y $ z $;
  COLUMN name age sex height weight;
  COMPUTE age;
   count+1;                                  /* 行の連番をセット */
   IF MOD(count,2)=1 THEN DO;                /* 奇数行のとき */
    CALL DEFINE(_ROW_, "STYLE", "STYLE=[BACKGROUND=Aliceblue]");
   END;
   ELSE DO;                                  /* 偶数行のとき */
    CALL DEFINE(_ROW_, "STYLE", "STYLE=[BACKGROUND=CXccffff]");
    END;
  ENDCOMP;
RUN;
```

Q IMPORTプロシジャでCSVファイルを読み込む際に、文字変数の値が切れてしまうことがあります。変数の長さを指定するなどで文字切れを回避することはできますか。

A **直接変数の長さを指定することはできませんが、最新のSAS9ではIMPORTプロシジャで新たに追加されたGUESSINGROWS=オプションで対応できます。このオプションで指定された範囲内のデータより、変数、データ型、データ長を判断します。オプション指定には事前に読み込む最大の行数を指定します。指定可能な値の範囲は1～32767までです。**

次の例では、GUESSINGROWS=オプションを使用して先頭から200行までのデータを事前に読み込み、データの判定を行なわせています。

例:GUESSINGROWS=オプションで先頭から200行を読み込む

```
PROC IMPORT OUT= WORK.test
  DATAFILE="C:¥temp¥test.csv"
  DBMS=CSV REPLACE;
  GETNAMES=YES;
  DATAROW=2;
  GUESSINGROWS=200 ;                /* 先頭から 200行を読み込む */
RUN;
```

なお、SAS8のIMPORTプロシジャは、標準ではCSVファイルの先頭の20行を走査して、変数の長さが決定されます。先頭の21行目以降に最大長のデータが存在する場合、変数の長さを判定させるには、次のような方法で対応可能です。

［対応1］SASレジストリ修正で最初に読み込む行数の設定値を変更する

1. SASを起動し、メニューの［ソリューション］>［アクセサリ］>［レジストリエディタ］を選択。
2. 左側のツリーを［PRODUCTS］>［BASE］>［EFI］の順に展開。
3. 右側の "GuessingRows" を選択して右クリックメニューから変更を選択。
4.「値のデータ」を20から適当な大きさに変更しOKを押す。たとえば、先頭50行を対象にしたい場合は50を入力します。
5.レジストリエディタを閉じる。ここで設定した行数をもとに、変数の長さが決定されるようになります。

［対応2］データの先頭行にダミーの最長データを挿入しておく

上記［対応1］の方法では、最長行が何行目にあるか予め把握しておく必要があります。また、あまりに大きな値を設定するとパフォーマンスが劣化する可能性も考えられます。このような場合、データの先頭行にあらかじめ最長データを挿入しておくことで文字切れを回避します。

Q データセットに含まれている変数の名前を変更したいと思っています。RENAMEステートメントを利用すれば可能なことは分かっているのですが、変更したい変数が多い場合、プログラムを記述するのが大変です。何か良い方法はありませんか。

A RENAMEステートメントの引数となる箇所を、あらかじめマクロ変数として定義しておくことで、変数名を列記する手間を省くことが可能です。RENAMEステートメントの引数は、"既存の変数名 = 新規の変数名" での指定となりますので、この指定部分を文字列としてあらかじめマクロ変数に格納しておきます。
次の例では、SQLプロシジャを用いて"既存の変数名=n_既存の変数名"とした文字列を変数の数だけ生成し、各々の文字列をブランク区切りでマクロ変数へ格納後にDATASETSプロシジャでのRENAMEステートメントの指定に利用しています。

例：SQLプロシジャでを利用したマクロ変数の生成

```
                                        /** テストデータの作成 **/
DATA a;
  col1=1;
  col3=3;
  col5=5;
  x=123;
RUN;
                                        /** マクロ変数の生成 **/
PROC SQL NOPRINT;
  SELECT TRIM(name)||'=n_'||TRIM(name) INTO:varlist separated by ' '
    FROM sashelp.vcolumn
    WHERE libname = "WORK" and
          memname = "A" and
          UPCASE(name) ? 'COL';
QUIT;
                                        /** 変数名の変更 **/
PROC DATASETS LIBRARY=work NOLIST;
  MODIFY a;
  RENAME &varlist;
QUIT;
                                        /** 変更の確認 **/
PROC CONTENTS DATA=a;
RUN;
```

上記プログラム例で生成されたマクロ変数には、以下の文字列が格納されます。

生成されるマクロ変数の内容例

```
col1=n_col1 col3=n_col3 col5=n_col5
```

Q SQLプロシジャのパススルー機能にてOracleのWHERE句にマクロ変数を利用したいと思っています。マクロ変数を展開させるためには、複引用符を使用しなければならないことは判っているのですが、OracleのWHERE句は構文上、単引用符を使用しなければなりません。何か良い方法はありませんか。

復引用符でマクロ変数を指定した例

```
WHERE ename="&MACV"
```

結果としてOracleの構文エラーが発生する

単引用符でマクロ変数を指定した例

```
WHERE ename='&MACV'
```

結果として不適切なWHERE句となる

A 特殊文字をクォートする%STRマクロ関数と%を利用して、単引用符をマークすることで対応可能です。次のプログラムを参考にしてください。

下記の例では、SQLパススルーのWHERE句に記述する条件式にて、変数名enameの値をマクロ変数&MACVとして定義できるように%STRマクロ関数と%でマクロ変数を定義しています。

例：パススルーSQLでのマクロ変数の使用

```
%LET macv=ALLEN;
PROC SQL;
  CONNECT TO ORACLE(USER=xxx PASSWORD=xxx PATH="@xxx");
  SELECT * FROM CONNECTION TO ORACLE
  (
    SELECT COUNT(*) FROM emp WHERE ename=%STR(%'&MACV%')
  );
  DISCONNECT FROM oracle;
QUIT;
```

Q 文字と数字が混在しているデータがあります。この中から、文字や数字を取り出す方法はありますか。

A これまではSCAN関数、SUBSTER関数などの利用で特定の文字を抽出することなどが可能でしたが、SAS9よりCOMPRESS関数に追加された機能を利用することで、簡易に文字・数字のみを取り出すことが可能となりました。
次の例ではCOMPRESS関数の3番目の引数に値を保持することを意味する "K "と、数値を表す "D "および文字（アンダーバーと英字）を表す "F "を組み合わせて指定し、変数内の数値と文字を取り出しています。

COMPRESS関数の構文

```
  COMPRESS(<source><, chars><, modifiers>)
```

説明

source　取り除きたい文字を含むSAS文字式
chars　SAS文字式から取り除きたい、1つ以上の文字
modifiers　COMPRESSの動作に対する設定

使用例

```
                                        /* テストデータ作成 */
DATA sample ;
  INPUT data1 $CHAR15. ;
DATALINES ;
2006 01 08 aaaaa
bbbb 2006-01-09
2006/01/10 cc
;
RUN ;

DATA ext ;
  SET sample ;
                                       /* KD 数値を残す */
  rc1 = COMPRESS(data1,,'KD') ;
                                       /* KF 文字を残す */
  rc2 = COMPRESS(data1,,'KF') ;
RUN ;

PROC PRINT DATA=ext (KEEP=rc1 rc2) ;
RUN ;
```

上記の使用例を実行すると、結果は以下のようになります。

OBS	rc1	rc2
1	20060108	aaaa
2	20060109	bbbb
3	20060110	cc

COMPRESS関数の詳細は、以下のURLやオンラインヘルプなどからご参照ください。

http://support.sas.com/onlinedoc/913/docMainpage.jsp
SAS OnlineDoc > Base SAS > SAS Language Reference: Dictionary > Dictionary of Language Elements > Functions and CALL Routines > COMPRESS Funtion

なお、COMPRESS関数の拡張は全角文字に対応していません。また、KCOMPRESS関数には機能の追加はありません。

現在実行しているSASプログラムのファイル名を取得する方法はありますか。

SASをバッチモードで実行している場合、SYSINオプションに実行ファイル名が格納されています。このオプションの値を参照することで、実現可能です。また、SAS9よりDMSモードにてSAS_EXECFILEPATH環境変数内に、実行ファイル名が格納されるようになりました。DMSモードで使用している場合は、この環境変数の値を参照することで実現可能です。

※SAS EXECFILEPATH環境変数は拡張エディタからプログラムを実行した場合のみ、参照可能です。

次の例では、SYSINオプションに指定されたファイルパスが無い場合に、%SYSGETマクロ関数を利用してSAS®EXECFILEPATH環境変数を取得するようにしています。

例:SAS_EXECFILEPATH環境変数の取得

```
%LET execpath=" ";
%MACRO setexecpath;
  %LET execpath=%SYSFUNC(GETOPTION(SYSIN));
  %IF %LENGTH(&execpath)=0
%THEN %LET execpath=%SYSGET(SAS_EXECFILEPATH);
%MEND setexecpath;

%setexecpath;
%PUT &execpath;
```

& や % を含む文字列を使ってマクロ変数を作成するには、どのようにすればよいですか。

マクロはプログラムテキストを ”トークン ”と呼ばれる単位に分解する ”ワードスキャナ ”を通してコンパイルされます。トークンはワードスキャナが他のトークンの先頭、トークン後のブランクを検出すると終了します。&や、%はこのトークンの1種類である特殊文字であり、コンパイラにとって意味のある文字です。このような、マクロにおいて構文的に意味の記号を単なるテキストとして扱うためには文字列を引用符で囲むクォート処理を行なう必要があります。クォート処理を行なうマクロ引用符関数はいくつかありますが、ここでは展開前の引数をクォートする%NRSTRマクロ引用符関数を紹介します。%STR 関数は以下の特殊記号をクォートします。

```
+ - * / < > = ¬ ^ ~ ; , blank
AND OR NOT EQ NE LE LT GE GT
```

また、対象となる記号の前に%記号をつけることにより、以下の記号もクォートします。

```
' " ( )
```

%NRSTR 関数は上記特殊記号に加え、& , % 記号についてもクォートします。%NRSTR 関数を使用したいくつかのサンプルパターンを記載しますので、参考にしてください。

例:%NRSTRマクロ引用符関数を使用したパターン例

```
%LET text1 = %NRSTR(M & A) ;           /*テキストの全てをクォート*/
%LET text2 = M %NRSTR(&) A ;            /* & 記号のみをクォート*/
%LET an    = %NRSTR(&)    ;          /*事前に &記号をクォート */
%LET text3 = M &an A       ;     /* 事前にクォートされた値を用いる*/
%PUT text1=&text1 text2=&text2 text3=&text3 ;   /* 結果確認 */
```

例を実行した結果は次のようになります。

```
text1=M & A text2=M & A text3=M & A
```

AUTOREGプロシジャを実行し、OUTEST=オプションを用いてパラメータ推定値をデータセットに出力しています。この際、作成されるデータセットに変数_STATUS_が含まれます。この変数は何を表しているのでしょうか。また、どのような値を取り得ますか。

変数_STATUS_は、AUTOREGプロシジャのモデル推定における反復過程において、収束しているかの情報を、0、1、2、もしくは3の値にて表しています。各値における解釈に関しては、以下の通りです。

0・・・ 収束基準を満たしています。
1・・・ 最適化の過程において、関数の値をより大きくすることができていません。
2・・・ 指定されている反復回数（MAXIT=50（デフォルト）にて、収束基準を満たしていません。（この場合、反復過程における最後の数値がパラメータの値として表示されます。）
3・・・ 上記以外のエラーが生じています。

変数_STATUS_が0以外の場合には、ログ、およびアウトプットにおけるWarning、Errorメッセージを確認してください。

多次元正規分布に従う乱数列を生成するにはどのようにしたらよいでしょうか。

幾つかの方法が考えられます。以下の例ではいずれも、3次元正規分布に従う乱数列を100オブザベーション作成しています。なお、平均ベクトル、および共分散行列として全て共通の値を使用しています。また、Mersenne-Twister（MT）による乱数生成は、SAS8.2では評価版の機能です。

1.VNORMAL Callを使用する

SAS8 以降では、SAS/IMLのVNORMAL Callを使用して生成することができます。なお、これによって生成される乱数列は旧来の乗算型合同法に基づくものです。次の例を参考にしてください。

例:VNORMAL Callの例

```
PROC IML;
  nobs=100;                          /** 生成するオブザベーション数 **/
  seed=12345;                              /** 乱数系列のシード **/
  mean={20 30 40};                            /** 平均の定義 **/

  cov={1    1.2  2.25,
       1.2  4    3.3,
       2.25 3.3  9};                       /** 共分散行列の定義 **/
                          /** 戻り値の乱数列、平均ベクトル、共分散行列、
                              オブザベーション数、シードの順番で指定**/
  CALL VNORMAL(rv,mean,cov,nobs,seed);
                                          /** SASデータセットへの出力
                                 データセット mnormal1 が作成されます **/
  CREATE mnormal1 FROM rv;
  APPEND FROM rv;
QUIT;
```

2. SAS/IMLでCholesky 分解に基づく処理を行なう。

Cholesky分解を行なうROOT関数を使用して、次のようなプログラムで生成することができます。なお、SAS9以降でサポートされているRANDSEED CallとRANDGEN Callを使用すると、MTに基づく乱数列を生成できます。次の例を参考にしてください。

例:SAS9以降のMTによる例

```
PROC IML;
  nobs=100;
  seed=12345;
  mean={20 30 40};
  cov={1    1.2  2.25,
       1.2  4    3.3,
       2.25 3.3  9};
  CALL RANDSEED(seed);                        /** シードの初期化 **/
                                  /** 100×3の行列を事前に作成しておく**/
  rvn=J(nobs,NCOL(cov),.);
  CALL RANDGEN(rvn,'NORMAL');                  /** 正規乱数の生成**/
                          /** Choleskey 分解を関数 ROOT で行なう **/
  rv=mean#J(nobs,NCOL(cov),1)+rvn*ROOT(cov);
                                          /** SASデータセットへの出力
                                 データセット mnormal2 が作成されます **/
  CREATE mnormal2 FROM rv;
  APPEND FROM rv;
QUIT;
```

例:乗算型合同法の例

```
PROC IML;
  nobs=100;
  seed=12345;
  mean={20 30 40};
  cov={1    1.2  2.25,
       1.2  4    3.3,
       2.25 3.3  9};
  rv=mean#J(nobs,NCOL(cov),1)
       + RANNOR(J(nobs,NCOL(cov),seed))*ROOT(cov);
                                          /** SASデータセットへの出力
                                 データセット mnormal3 が作成されます **/
  CREATE mnormal3 FROM rv;
  APPEND FROM rv;
QUIT;
```

3.SAS/ETSのMODELプロシジャの利用

SAS/ETSのMODELプロシジャには、シミュレーションを行なう機能が備わっています。これを利用して、MT、および乗算型合同法による乱数列を生成することができます。PSEUDO= オプションは、SAS9 以降でのみサポートされており、SAS8では利用できません。

例：SAS9以降のMODELプロシジャによるMTの例

```
                    /** 平均ベクトルからなる SAS データセットを作成 **/
DATA mean1;
  _NAME_="";
  INPUT a1-a3;
DATALINES;
20 30 40
;
RUN;
                    /** 共分散行列からなる SAS データセットを作成 **/
DATA cov1;
  INPUT _NAME_$ col1-col3;
DATALINES;
col1 1    1.2 2.25
col2 1.2  4   3.3
col3 2.25 3.3 9
;
RUN;
                                        /** MODELプロシジャ **/
PROC MODEL DATA=_NULL_ NOPRINT;
  PARMS a1 a2 a3;
  col1=a1;
  col2=a2;
  col3=a3;
                  /** SOLVEステートメントでは、ESTDATA= で平均ベクトル、
                   SDATA= で共分散行列からなるデータセット名を指定します。
           また、RANDOM=でオブザベーション数を、SEED=でシードを与えます **/
                     /** SAS9.1では、PSEUDO=TWISTERを指定するとMT、
                      PSEUDO=DEFAULTと指定すると乗算型合同法に基づく
                  乱数列を生成します。SAS8では、このオプションは利用できず、
                              乗算型合同法に基づいた方法のみ利用できます **/
  SOLVE col1-col3/ESTDATA=mean1 SDATA=cov1 RANDOM=100 SEED=12345
      PSEUDO=TWISTER
      OUT=mnormal4(WHERE=(_REP_^=0) DROP=_TYPE_ _MODE_ _ERRORS_);
RUN;
QUIT;
```

4.Cholesky分解をMIXEDプロシジャで行なう

Cholesky分解はMIXEDプロシジャの内部でも行なわれており、この機能を応用して実現することも可能です。

例：SAS9以降のMTによる例

```
%LET nobs=100;                          /** オブザベーション数 **/
%LET ncol=3;                                  /** 次元 **/
%LET seed=12345;                              /** シード **/
%LET out=mnormal5;                      /** 出力データセット名 **/
                /** 平均と共分散行列からなる SAS データセットを作成 **/
                                      /** 平均は、「縦」に入れる **/
DATA cov2;
  INPUT col1-col3 mean;
  row+1;
DATALINES;
 1    1.2 2.25 20
 1.2  4   3.3  30
 2.25 3.3 9    40
;
RUN;
                  /** 標準正規分布に従う乱数を100×3 だけ作成する **/
                     /** 関数 RAND ではなく関数 RANNOR を使用すると、
                      乗算型合同法に基づく乱数列を生成することになる**/
DATA _random;
  CALL STREAMINIT(&seed);
  _TYPE_="SCORE";
  _MODEL_="col";
  mean=1;
  ARRAY col{&ncol};
  DO num=1 TO &nobs;
    DO i=1 TO DIM(col);
      col{i}=RAND("NORMAL");
    END;
    output;
  END;
  DROP i;
RUN;
```

例：MIXEDプロシジャを利用した例

```
        /** MIXEDプロシジャは、Cholesky 分解を行なう目的のためにのみ使用
            しており、プログラムそのものには統計的な意味づけはできません **/
ODS LISTING CLOSE;
ODS OUTPUT CHOLG=_Cholesky;
PROC MIXED DATA=cov2;
  CLASS Row mean;
  PARMS /NOITER;
  MODEL mean=;
  RANDOM row*mean/TYPE=UN GDATA=cov2 GC;
RUN;
ODS LISTING;
                  /** SCORE プロシジャで行列の乗算を擬似的に行なう **/
PROC SCORE DATA=_cholesky SCORE=_random OUT=_out(KEEP=col num);
  BY num;
  VAR mean col:;
RUN;

PROC TRANSPOSE DATA=_out OUT=&out.(DROP=_NAME_);
  BY num;
RUN;
```

New Publications
新刊マニュアルのお知らせ

「The Little SAS® Book for Enterprise Guide® 3.0」
注文番号:59376
I S B N:0-471-75451-X
価　　格:9,660円(税込)

「The Complete Guide to SAS® Indexes」
注文番号:60409
I S B N:1-59047-849-5
価　　格:10,605円(税込)

「Information Revolution: Using the Information Evolution Model to Grow Your Business」
注文番号:60887
I S B N:0-471-77072-8
価　　格:3,780円(税込)

SAS Learning Edition リリース 2.0値下げのご案内

発売以来、多くの方にご利用いただいているSAS Learning Edition リリース 2.0を2006年3月1日より値下げいたしました。

一 般 利 用	36,750円(税込)
アカデミック利用	26,250円(税込)

詳しくは弊社ホームページ(http://www.sas.com/japan/manual/le.html)をご覧ください。

マニュアルパッケージのご案内

「マニュアルを購入したいが、どのマニュアルを購入すればいいかわからない」、「必要なマニュアルをまとめて購入したい」など、このようなお客様の声にお応えし、この度、お得なマニュアルパッケージの販売を開始しました。用途に即したSASマニュアルを、お手元に置いてぜひご活用ください。

A:医薬向けパッケージ
B:経済向けパッケージ
C:統計解析基礎パッケージ
D:中～上級者向け統計解析パッケージ
E:統計解析ベストパッケージ
　「C:統計解析基礎パッケージ」＋「D:中～上級者向け統計解析パッケージ」の内容です。
F:パーフェクトパッケージ
　「E:統計解析ベストパッケージ」に、さらに役立つ詳細なマニュアルを追加した内容です。

お申し込みについては専用注文用紙にて承ります。詳しくは弊社ホームページ(http://www.sas.com/japan/manual/package.html)をご覧ください。SASマニュアル申込用紙、および最新のPublication Catalog(マニュアル案内パンフレット)は弊社ホームページ(http://www.sas.com/japan/manual/)にて公開しておりますので、併せてご利用ください。

マニュアル販売係

- T E L 03-3533-3835
- F A X 03-3533-3781
- E-mail JPNBooksale@sas.com

SAS Training
SASトレーニングのお知らせ

■ 特別トレーニングコース開催のご案内

●「SASによる遺伝子多型データの解析入門」コース

日　　程: 東京会場:2006年4月20日(木) 10:00～17:00
価　　格: 50,400円(税込)／チケット捺印数 1
会　　場: 東京会場:SAS Institute Japan株式会社 東京本社 7Fトレーニングルーム
受講対象: ゲノムデータ解析を担当している方、またはこれから担当予定の方
前提知識: 「医薬向けカテゴリカルデータ解析2」を受講済みか、同程度の知識のある方
学習内容: 近年医学研究において、遺伝子と疾患、遺伝子と環境因子と疾患との関連を調べるために、遺伝子多型データが頻繁に用いられるようになりました。他の共変量と異なり、多型データはその変数が得られた遺伝学的背景を考慮した解析手法を適用する必要性が生じます。そこで、本コースでは、多型データに対する一連の基礎的な解析方法について、主にSAS/Geneticsを用いて解説します。同時にSAS/STATの適当なプロシジャでの実行例や、近年の大量タイピングデータに対応するためのSAS Enterprise Minerを用いた方法なども示します。

主に解説する方法
- アリル頻度の推定
- HW平衡検定
- 連鎖不平衡検定・連鎖不平衡値の推定
- ハプロタイプの推定
- 関連分析の基礎
- 決定木を用いた探索的方法 など

なお、遺伝学の基礎的な概念や用語に関してはできる限りの解説を行ないますが、生物学の基礎知識程度はあることが望まれます。

■ 新しいディスカウント制度導入のご案内

ディスカウント制度に新しく「グループ受講割引」および「SAS認定試験対策割引」が追加になりました。

● グループ受講割引

弊社ホームページ上にて公開されているトレーニングを、同一企業で同一コースをまとめて3名様以上でお申し込みいただくと、通常価格から10%割引でご受講いただけます。
グループ受講割引の詳細については以下のURLをご参照ください。
http://www.sas.com/japan/training/group.html

● SAS認定試験対策割引

SAS認定試験の受験に必要な各種トレーニングコースを、セットにして割引料金でご受講いただけるサービスです。

プラン名／内容	税込価格
■SAS Certified Base Programmer 試験対策プラン(合計2コース) ・「SASプログラミング I」 ・「SASプログラミング II」 ※ 認定試験料の20%割引券付き ※ 受講有効期間:最初に受講したコースの初日から3ヶ月間	通常291,375円 ↓ 233,100円 (20%割引)
■SAS Certified Advanced Programmer 試験対策プラン(合計3コース) ・「SASプログラミング III」 ・「SASによるSQL入門」 ・「SASマクロ言語入門」 ※ 認定試験料の20%割引券付き ※ 受講有効期間:最初に受講したコースの初日から6ヶ月間	通常349,125円 ↓ 279,300円 (20%割引)

SAS認定試験対策割引の詳細については以下のURLをご参照ください。
http://www.sas.com/japan/training/certset.html

■ 2006 年度版トレーニングカタログのご案内

ただいま2006年度版トレーニングカタログをご希望のお客様へ郵送にてお送りするサービス(無料)を行なっております。ご希望のお客様は、住所、会社名、部署名、氏名を必ずご記入の上、弊社トレーニング担当宛にE-mailにてご連絡ください。

SAS Institute Japan株式会社では、今後も多岐にわたったトレーニングコースを追加していく予定です。コース内容・日程等の詳細は、順次弊社Webサイトに公開しますので、以下のURLをご参照ください。
http://www.sas.com/japan/training/

その他、トレーニングに関する情報については、上記のURLをご参照いただくか、下記トレーニング担当までお問い合わせください。

トレーニング担当

● T E L 03-3533-3835
● F A X 03-3533-3781
● E-mail JPNTraining@sas.com

Latest Releases
最新リリース情報

■PCプラットフォーム

Windows版	SAS 9.1.3	9.1 TS1M3
64-bit Windows (Itanium)版	SAS 9.1.3	9.1 TS1M3

■UNIXプラットフォーム

Tru64版	SAS 9.1.3	9.1 TS1M3
SunOS/Solaris版	SAS 9.1.3	9.1 TS1M3
HP-UX版	SAS 9.1.3	9.1 TS1M3
HP-UX (Itanium)版	SAS 9.1.3	9.1 TS1M3
AIX版	SAS 9.1.3	9.1 TS1M3
Linux (Intel)版	SAS 9.1.3	9.1 TS1M3
ABI+版	SAS 6.11	TS040

■ミニコンピュータプラットフォーム

OpenVMS AXP版	SAS 6.12	TS020
OpenVMS VAX版	SAS 6.08	TS407

■メインフレームプラットフォーム

IBM版(OS/390, z/OS)	SAS 9.1.3	9.1 TS1M3
富士通版(F4, MSP)	SAS 6.09E	TS470
日立版(VOS3)	SAS 6.09E	TS470
CMS版	SAS 6.08	TS410

Information
SAS Technical News 送付についてのご案内

SAS Technical Newsは次の方を対象にお送りしています。

- SASコンサルタントとしてご登録の方
- SAS Technical Newsの購読をお申し込みいただいている方

今後SAS Technical News購読が不要の方、配信先の変更等をご希望の方は、下記URLよりお手続きください。

配信停止
http://www.sas.com/japan/corporate/material.html

配信先変更手続き
http://www.sas.com/japan/sasj_privacy.html#inquiry

SAS Technical News Spring 2006
発行
SAS Institute Japan株式会社

テクニカルニュースに関するお問い合わせ先
テクニカルサポートグループ
TEL:03-3533-3877
FAX:03-3533-3781
E-mail:JPNTechnews@sas.com

東京本社
〒104-0054
東京都中央区勝どき1-13-1
イヌイビル・カチドキ
Tel 03 (3533) 6921
Fax 03 (3533) 6927

大阪支店
〒530-0004
大阪市北区堂島浜1-4-16
アクア堂島西館 12F
Tel 06 (6345) 5700
Fax 06 (6345) 5655

SAS Institute Japan株式会社 www.sas.com/japan/