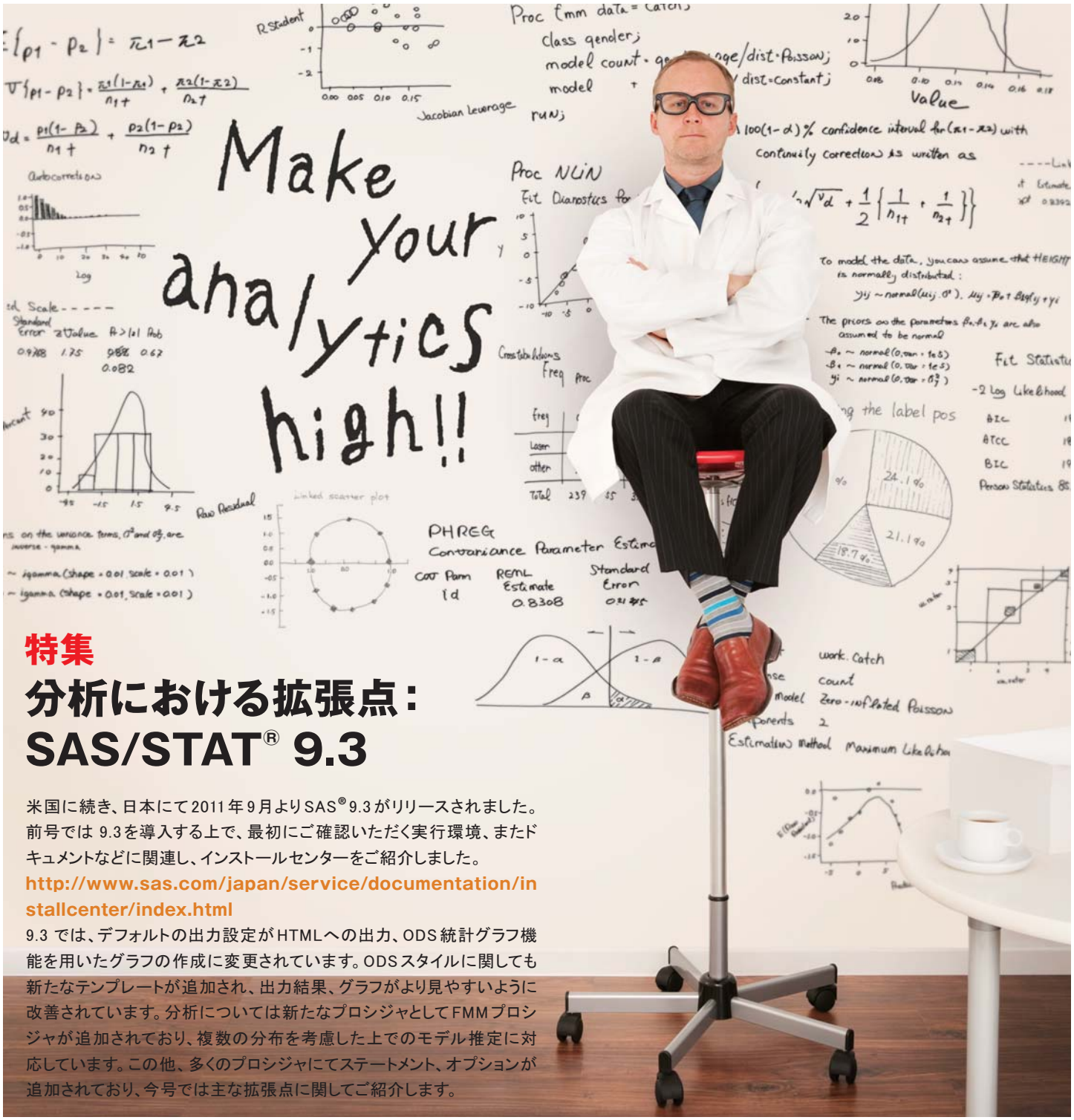
STN
SAS Technical News

For Higher
Customer Satisfaction,
We Bridge
the SAS System
Between
Customer's World.

AUTUMN 2011



特集
分析における拡張点：SAS/STAT® 9.3

米国に続き、日本にて2011年9月よりSAS®9.3がリリースされました。前号では 9.3を導入する上で、最初にご確認いただく実行環境、またドキュメントなどに関連し、インストールセンターをご紹介しました。

http://www.sas.com/japan/service/documentation/installcenter/index.html

9.3 では、デフォルトの出力設定がHTMLへの出力、ODS 統計グラフ機能を用いたグラフの作成に変更されています。ODSスタイルに関しても新たなテンプレートが追加され、出力結果、グラフがより見やすいように改善されています。分析については新たなプロシジャとしてFMMプロシジャが追加されており、複数の分布を考慮した上でのモデル推定に対応しています。この他、多くのプロシジャにてステートメント、オプションが追加されており、今号では主な拡張点に関してご紹介します。

1 SAS 9.3におけるデフォルトの出力

SASを起動した際、入門ガイドに関するウィンドウが表示されます。9.3では入門ガイドのご案内に加え、出力に関するデフォルトの設定が変更されていることが表示されます。

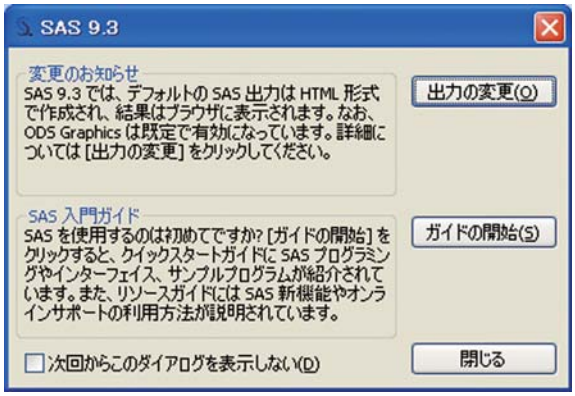


図1：起動時に表示されるウィンドウ

まず最初に、このデフォルト設定の変更について説明します。

1.1 デフォルト設定に関して

以前のリリースでは、出力結果はアウトプットウィンドウに出力されます。LISTING出力と言及され、この部分の出力を抑制するためには、ODS LISTING CLOSE; ステートメントを記述します。9.3では、デフォルトの設定はHTML出力に変更されています。また、9.1.3では評価版、9.2では正規の機能となりましたODS統計グラフ機能に関し、デフォルトにて機能が有効となっています。このため、以前のリリースとは異なり、ODS GRAPHICS ON;ステートメントを記述する必要がなくなりました。
HTML出力、ODS統計グラフ機能が有効であるデフォルトの設定では、ODS統計グラフは分析結果と併せ、一つのHTMLファイルに表示されます。この変更にともない、出力のイメージを設定しているODSスタイルに関しても新たなスタイルが追加されており、そのうちの一つHTMLBLUEがデフォルトの設定となっています。（以前のリリースにおけるHTML出力では、ODSスタイルは“Default”でした。）

1.2 デフォルト設定の変更

以前のリリースとの比較、また使い慣れている環境と同じにしたいなど、デフォルトの設定を変更したい場合もあります。この場合、SASのメニューより以下の選択を行います。

[ツール]→[オプション]→[プリファレンス]

図2：プリファレンスウィンドウ

図2の結果タブがデフォルトの設定となっており、この画面にて設定の変更を行えます。例えば、HTML出力を抑制し、アウトプットウィンドウへの出力を行う場合には、“リストを作成する”を選択し、“HTMLを作成する”のチェックマークを外します。また、スタイルに関しては、▼の部分をクリックすることによって、リストより選択できます。

1.3 ドキュメントに関して

この変更点に関するドキュメントについては、SASを起動した際に表示されるウィンドウ（図1）で“出力の変更”を選択することで参照できます。また、ODS機能の新機能に関する、以下のドキュメントの箇所にも記載があります。

http://support.sas.com/documentation/cdl/en/whatsnew/64209/HTML/default/viewer.htm#odsugwhatsnew93.htm

2 SAS/STATにおける新プロシジャ：FMMプロシジャ（評価版）

SAS 9.2ではGENMODプロシジャにてZEROMODELステートメントが追加され、ゼロ強調モデル（Zero Inflated Model）の推定ができるようになりました。このモデルは、2つの分布、Poisson分布（もしくは負の二項分布）と退化分布（degenerate distribution）に基づいています。SAS 9.3では、これらのモデルを包括する、有限混合モデルに対し、FMMプロシジャが追加されています。

2.1 有限混合モデルのモデル式

Yes、Noなどの2値データに対する二項分布、カウントデータに対するPoisson分布などを用いたモデルは、一般化線形モデル（Generalized Linear Model）と言及され、LOGISTICプロシジャ、GENMODプロシジャ、GLIMMIXプロシジャなどが対応しています。これらのモデルでは、平均と分散が互いに依存するため、実際のデータに対し、分析を行った場合、過分散（overdispersion）が生じ、検定結果が正しくないなどの問題が生じるケースがあります。
過分散が生じる一つの例としては、カウントデータにおいて0の値が多く含まれていることが挙げられます。観測した際の値がたまたま0であったこと以外に、必然的に値0であったなど、異なる起因にて生じていることがあります。このようなケースにおけるモデルの一つとして、ゼロ強調モデルがあります。カウントに対する分布をPoisson分布と仮定した場合、このモデルは以下のように記述できます。

モデル1

$$Y \mid S \sim \begin{cases} Poisson(\lambda) & \text{if } S=1 \\ 0 & \text{if } S=0 \end{cases}$$

変数Sの分布は二項分布（試行数1、確率p）であると仮定します。多くの場合、変数Sの値は観測されないため、潜在変数ともいわれます。Poisson分布のパラメータλ（平均）を4、二項分布のパラメータp（確率）を0.9とした場合、疑似的なサンプルに基づく、ヒストグラムは以下のようになります。

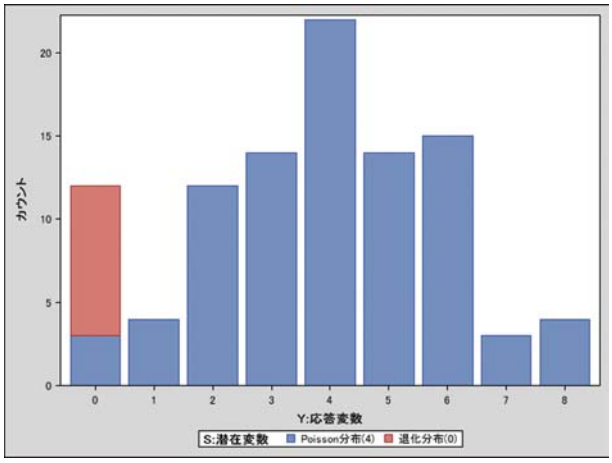


図3：シミュレーション結果に基づくヒストグラム

Y=0で観測されているカウントは12となりますが、Poisson分布に基づくカウントは3となっており、そのほか（9）は退化分布に基づき、0として観測されています。つまり、2つの分布が混合しているため、一つの分布、Poisson分布のみを用いて分析を行った場合、推定結果が正しくなく、検定結果に基づく判断に誤りが生じてしまうかもしれません。
Poisson分布を用いたゼロ強調モデルとして、（モデル1）の式を記述しましたが、より一般的には密度関数を用います。Poisson分布、退化分布をそれぞれ $f_1(y|S=1)$、$f_2(y|S=0)$ とし、二項分布をPr(S=j)とした場合、以下のように記述できます。

$$f(y)=\Pr(S=1)\times f_1(y\mid S=1)+\Pr(S=0)\times f_2(y\mid S=0)$$

簡略な記述とするため、2つの分布を混合したモデル、ゼロ強調モデルをここまでご紹介してきました。新たなFMMプロシジャでは、有限であればより多くの分布を混合したモデルの推定ができ、より一般的なモデル式としては、以下のようになります。

モデル2

$$f(y;\alpha,\beta)=\Sigma_{j=1}^{k}\Pr(S=j)\times f_j(y;\alpha_j,\beta_j\mid S=j)$$

各分布f_jに対して説明変数、分布の指定が必要となるため、FMMプロシジャでは複数のMODELステートメントを記述します。分布に関しては、DIST=オプションを用い、二項分布（BINOMIAL）、Poisson分布（Poisson）、ガンマ分布（GAMMA）、正規分布（Gaussian）、退化分布（CONSTANT）などを指定できます。リンク関数については、LINK=オプションを用います。また、潜在的な変数Sの分布Pr(S=j)に関してはPROBMODELステートメントにて指定することができます。

2.2 他プロシジャとの関連

前節のモデル式にあるように、有限混合モデルは複数の一般化線形モデルに基づいています。このため、一つのみの一般化線形モデルにも対応しており、モデルの推定ができます。つまり、CATMODプロシジャ、GLMプロシジャ、LOGISTICプロシジャ、GENMODプロシジャ、GLIMMIXプロシジャ、REGプロシジャにて対応しているモデルに対しても推定が行えます。
モデルの推定を行った後、カテゴリカル変数の各水準に対するLS平均、LS平均の差（LSMEANSステートメント）、また、推定可能であるパラメータ線形式の推定（ESTIMATEステートメント）、対比に対する検定（CONTRASTステートメント）が必要となる場合があります。しかしながら、線形モデルにおける推定可能性などの理論が直接、有限混合モデルに適用することができないため、FMMプロシジャではこれらのステートメントはサポートされていません。

2.3 例題：ゼロ強調 Poisson 回帰モデル

FMMプロシジャを用いた分析の一例として、ゼロ強調 Poisson 回帰モデルを用います。
サンプルデータは、公園で過去 6 カ月に釣った魚の数をアンケートした結果です。この際、性別と年齢は質問していますが、釣りを行ったかの情報は含まれていません。

サンプルテーブル

```
DATA catch;
   INPUT gender $ age count @@;
   DATALINES;
  F  54 18  M  37   0  F  48 12  M  27   0  M  55   0
  M  32  0  F  49 12  F  45 11  M  39   0  F  34   1
  F  50  0  M  52   4  M  33  0  M  32   0  F  23   1
  F  17  0  F  44   5  M  44  0  F  26   0  F  30   0
  F  38  0  F  38   0  F  52 18  M  23   1  F  23   0
  M  32  0  F  33   3  M  26  0  F  46   8  M  45   5
  M  51 10  F  48   5  F  31  2  F  25   1  M  22   0
  M  41  0  M  19   0  M  23  0  M  31   1  M  17   0
  F  21  0  F  44   7  M  28  0  M  47   3  M  23   0
  F  29  3  F  24   0  M  34  1  F  19   0  F  35   2
  M  39  0  M  43   6
;
```

このサンプルデータからは、観測されているカウントとして 0 が比較的多く含まれているようです。魚の数はカウントデータであるため、最初に Poisson 回帰モデルを推定します。GENMOD プロシジャにて行えますが、次のように FMM プロシジャにても推定できます。

```
PROC FMM DATA=catch;
  CLASS gender;
  MODEL count = gender*age / DIST=Poisson;
RUN;
```

他のプロシジャのように、カテゴリカルな説明変数を含む場合には CLASS ステートメントを用います。また、MODEL ステートメントにて応答変数（COUNT）、効果として GENDER*AGE を記述した上で、DIST=POISSON オプションを追記し、Poisson 分布を指定しています。変数 GENDER はカテゴリカル変数（CLASS ステートメントにても指定）となるため、効果 GENDER*AGE の指定にて、変数 GENDER の水準ごとに変数 AGE に対する異なる傾きをモデルに含めていることになります。実行結果として、以下のように適合度統計量（Fit Statistics）、パラメータ推定値などが出力されます。
過分散などの問題が生じていない場合、Pearson Statistic はおおよそ（サンプル数 − パラメータ数）となります。このデータセットにおけるサンプル数は 52、パラメータの数は 3 であることより、Pearson Statistics の値はおおよそ 49 となりますが、実際には 85.9573 と算出されており、このモデルでは過分散の問題が生じているのではないかと推察されます。

Fit Statistics	
−2 Log Likelihood	182.7
AIC (smaller is better)	188.7
AICC (smaller is better)	189.2
BIC (smaller is better)	194.6
Pearson Statistic	85.9573

Parameter Estimates for 'Poisson' Model

Effect	gender	Estimate	Standard Error	z Value	Pr > |z|
Intercept		−3.9811	0.5439	−7.32	<.0001
age*gender	F	0.1278	0.01149	11.12	<.0001
age*gender	M	0.1044	0.01224	8.53	<.0001

図4：Poisson回帰モデルの出力結果(一部)

また、このアンケートでは過去に釣った魚の数のみを聞いているため、0 と答えた人には釣りをまったくしなかった人と、釣りをしたが残念ながら一匹も釣れなかった人がいると考えらえます。つまり、2 つのグループ（分布）が混在していると思われるため、前述の FMM プロシジャのプログラムに、もう一つの分布に対する MODEL ステートメントを追加した上で、再度、実行します。

```
PROC FMM DATA=catch;
  CLASS gender;
  MODEL count = gender*age / DIST=Poisson ;
  MODEL       +            / DIST=Constant;
RUN;
```

2 つ目の MODEL ステートメントは ' + ' のみとなっています。この場合、前述の MODEL ステートメントにおける変数の指定を引き継いでおり、応答変数を繰り返し、記述する必要がありません。DIST=CONSTANT オプションと追記することによって、退化分布（デフォルトでは 0）を指定しています。
出力結果の最初の部分では、モデルに関する情報 'Model Information' が表示されます。項目の一つにはモデルのタイプ（Type of Model）があり、Zero-inflated Poisson と表示されていますことをご確認ください。

Model Information	
Data Set	WORK.CATCH
Response Variable	count
Type of Model	Zero-inflated Poisson
Components	2
Estimation Method	Maximum Likelihood

図5-1：ゼロ強調Poisson回帰モデルの出力結果(一部)

適合度統計量の部分では、Poisson 回帰モデルの時（図 4）と比べ、値が小さくなっており、より適したモデルとしてとらえられます。また、Pearson Statistic の値もほぼ半分になっており、より適したモデルであることが示唆されます。

Fit Statistics	
-2 Log Likelihood	145.6
AIC (smaller is better)	153.6
AICC (smaller is better)	154.5
BIC (smaller is better)	161.4
Pearson Statistic	43.4467
Effective Parameters	4
Effective Components	2

Parameter Estimates for 'Poisson' Model

Component	Effect	gender	Estimate	Standard Error	z Value	Pr > |z|
1	Intercept		-3.5215	0.6448	-5.46	<.0001
1	age*gender	F	0.1216	0.01344	9.04	<.0001
1	age*gender	M	0.1056	0.01394	7.58	<.0001

図5-2 : ゼロ強調Poisson回帰モデルの出力結果(一部)

最後にPoisson分布と退化分布における確率(Mixing Probabilities)が出力されており、69.72%がPoisson回帰モデルに依存していると解釈できます。

Parameter Estimates for Mixing Probabilities

Effect	Linked Scale				Probability
	Estimate	Standard Error	z Value	Pr > |z|	
Intercept	0.8342	0.4768	1.75	0.0802	0.6972

図5-3 :ゼロ強調Poisson回帰モデルの出力結果(一部)

ここまでは、モデルの推定を行う際、最尤法を用いています。図5-1における'Model Information' 表を再度参照しますと、最後の項目として推定方法があり、最尤法(Maximum Likelihood)と表示されています。この他、FMMプロシジャではベイズ分析などで用いられるMCMC(Markov Chain Monte Carlo)法による推定を行うことができます。この場合、前述のプログラムにBAYESステートメントを追加します。

```
PROC FMM DATA=catch;
  CLASS gender;
  MODEL count = gender*age / DIST=Poisson ;
  MODEL       +            / DIST=Constant;
  BAYES;
RUN;
```

出力結果における最初の表 'Model Information' において、推定方法として 'Markov Chain Monte Carlo' と表示され、乱数生成における初期シード値が併せて出力されます。また、線形モデルにおける各パラメータに対する事前分布に関する表が出力されます。ここでは、デフォルトで平均0、分散1000の正規分布となっており、2つのモデルの比率に関しては、Dirichlet分布が事前分布となっています。

Prior Distributions

Component	Effect	gender	Distribution	Mean	Variance	Initial Value
1	Intercept		Normal(0, 1000)	0	1000.00	-3.5215
1	age*gender	F	Normal(0, 1000)	0	1000.00	0.1216
1	age*gender	M	Normal(0, 1000)	0	1000.00	0.1056
1	Probability		Dirichlet(1, 1)	0.5000	0.08333	0.6972

図6-1 : MCMC法を用いた場合の出力結果(一部)

MCMC法を用いた場合、分析の結果として事後分布に関する情報が出力されます。

Posterior Summaries

Component	Effect	gender	N	Mean	Standard Deviation	Percentiles		
						25%	50%	75%
1	Intercept		10000	-3.5304	0.6345	-3.9601	-3.5153	-3.0932
1	age*gender	F	10000	0.1216	0.0133	0.1124	0.1213	0.1304
1	age*gender	M	10000	0.1054	0.0138	0.0958	0.1052	0.1148
1	Probability		10000	0.6921	0.0925	0.6280	0.6941	0.7572

Posterior Intervals

Component	Effect	gender	Alpha	Equal-Tail Interval		HPD Interval	
1	Intercept		0.050	-4.7576	-2.3191	-4.7572	-2.3190
1	age*gender	F	0.050	0.0961	0.1475	0.0975	0.1484
1	age*gender	M	0.050	0.0782	0.1329	0.0781	0.1324
1	Probability		0.050	0.5038	0.8694	0.5000	0.8520

図6-2 : MCMC法を用いた場合の出力結果(一部)

また、ODS統計グラフ機能を用いて、各パラメータのサンプルに関して、Trace Plotが作成されます。ここでは2つの分布における確率(Mixing Probability)に関するグラフを表示しています。

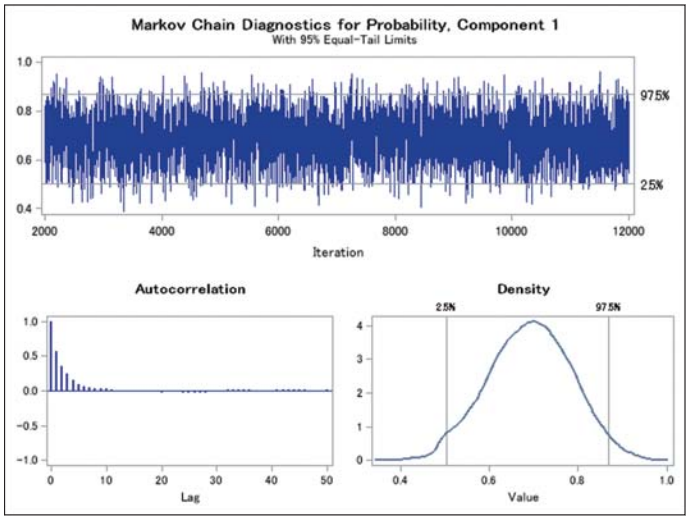


図6-3:MCMC法を用いた場合の出力結果(一部)

2つの分布を混合したゼロ強調Poisson回帰モデルを例題とし、FMMプロシジャのプログラム、および出力結果をご紹介しました。この例題に関する詳細に関しては、ドキュメントにてご参照いただくことができます。[SAS/STAT]→[SAS/STAT User's Guide]→[The FMM Procedure]→[Getting Started]→[Modeling Zero-Inflation: Is it Better to Fish Poorly or Not to Have Fished At All?]

また、以下の文献にても同じ例題を用いた記載があります。

On Deck: SAS/STAT® 9.3
http://support.sas.com/resources/papers/proceedings11/331-2011.pdf

3 SAS/STATにおける主な拡張点

有限混合モデルに対するFMMプロシジャが追加された他、多くのプロシジャにてステートメント、オプションが追加され、機能が拡張されています。ここでは、主な拡張点をご紹介します。

3.1 EFFECTステートメントに関して

9.22では評価版の機能であったEFFECTステートメントが正規版の機能として以下のプロシジャ(11)にてサポートされています。

HPMIXED	GLIMMIX	GLMSELECT
LOGISTIC	ORTHOREG	PHREG
PLS	QUANTREG	ROBUSTREG
SURVEYLOGISTIC	SURVEYPHREG	

説明変数をモデルに含める場合、MODELステートメントにて変数を記述します。カテゴリカルな変数に関しては、CLASSステートメントにて変数を追記することで、内部的にダミー変数を作成し、モデルの推定、分析が行われます。しかしながら、この他の変換が必要である場合、事前にDATAステップにて新たな変数を作成しておくなどの処理が必要でした。新たに正規版となりましたEFFECTステートメントでは、スプライン関数の変数などをモデルに含めることができます。
EFFECTステートメントでは、最初に効果名を指定し、等号の後にどのような効果であるかを記述します。効果名は、そのあとに指定するMODELステートメントにて記述します。

例

```
PROC GLIMMIX DATA=one;
  CLASS a b sub;
  EFFECT spl = SPLINE(x);
  MODEL y= a b spl a*spl;
  RANDOM a*b / SUBJECT=sub;
RUN;
```

EFFECTステートメントにて指定できる効果は以下の5つとなります。

● COLLECTION

複数の変数を一つの効果として扱い、モデルの推定、検定を行います。例えば、変数X1、X2にて一つの効果を表している場合、COLLECTION(X1 X2)の指定を行うことによって、変数ごとの検定ではなく、2つの変数を一つにまとめた効果に対する検定結果が出力されます。すでにダミー変数に展開されている効果などがある場合に有用となります。

例

```
EFFECT test = COLLECTION(x1 x2);
```

● LAG(評価版)

一時点前の水準を表すダミー変数を作成します。クロスオーバー分析などで、キャリーオーバー（持越し）効果を検証する際には、前の時点の投与薬を示す変数が必要となり、この指定が有用となります。なお、被験者を表す変数(subject)、時点を表す変数(period)の指定が必要となります。

例

```
CLASS treatment;
EFFECT lag = LAG(treatment / WITHIN=subject PERIOD=period);
```

● MULTIMEMBER | MM

複数のカテゴリカル変数から構成されるダミー変数を作成することができます。通常、オブザベーションごとに一つのダミー変数が1となり、他のダミー変数は0となりますが、このMULTIMEMBERを用いることで、複数のダミー変数を1とすることができます。

例

```
CLASS teacher1 teacher2;
EFFECT teacher = MULTIMEMBER(teacher1 teacher2);
```

● POLYNOMIAL | POLY

複数の連続変数に基づく、多項式の変数を作成します。

例

```
EFFECT  mypol = POLYNOMIAL(x1-x3 / degree=2);
```

● SPLINE

連続変数に対してスプライン展開を行い、複数のスプライン関数の和として表します。

例

```
EFFECT spl = SPLINE(x);
```

各効果の指定に関する詳細、オプションに関しては、ドキュメントの以下の箇所をご参照ください。
[SAS/STAT]→[Shared Concepts and Topics]→[EFFECT Statement]

3.2 ODS統計グラフ対応のプロシジャ

すでに多くのプロシジャがODS統計グラフ機能に対応していますが、新たに以下のプロシジャ(9)にてもODS統計グラフを作成できます。

FMM	GLMPOWER	NLIN
ORTHOREG	POWER	SURVEYLOGISTIC
SURVEYPHREG	SURVEYREG	VARCLUS

この他、CLUSTERプロシジャのデンドログラム(樹形図)、FREQプロシジャにおける一致統計量(AGREEオプション)のグラフが新たに追加されています。

また、ODS統計グラフ機能に関するオプションとしてMAXPOINTS=オプションが追加されています。GLMプロシジャなどを実行すると箱ひげ図、散布図などのグラフがデフォルトにて表示されますが、オブザベーションの数が非常に大きい場合には表示されません。この場合、MAXPOINTS=オプションにて上限を変更、もしくはMAXPOINTS=NONEの指定にて上限を設定しないことにより、グラフが表示できます。MAXPOINTS=オプションは以下のプロシジャにて対応しています。

ANOVA	CLUSTER	GLM
LOESS	LOGISTIC	MIXED
QUANTREG	REG	VARCLUS

※ MAXPOINTS=オプションのデフォルトは5000となります。ただし、CLUSTERプロシジャ、VARCLUSプロシジャに関しては、デフォルト値が200となっており、MAXPOINTS=NONEの指定はできません。

3.3 その他の主な拡張点

● CALISプロシジャ

COSANモデルにおける平均構造分析、RAMモデル指定の向上、推定方法としての全情報最尤法(FIML)など、9.22にて評価版として追加された機能が9.3では正規版となりました。また、FIML法を用いた場合、欠損パターンの詳細分析が提供されます。

● GLMSELECTプロシジャ

STOREステートメントが新たにサポートされ、選択されたモデルの情報をアイテムストアとして保存できます。保存したアイテムストアは、PLMプロシジャにて呼び出し、分析結果を求めることができます。

● HPMIXEDプロシジャ

反復測定分析などに用いるREPEATEDステートメントが追加されており、誤差項に対する共分散構造を指定できます。

● MCMCプロシジャ

ランダム効果の指定に関し、新たにRANDOMステートメントが追加されています。このため、階層的なモデル(Hierarchical Model)などランダム効果に対するプログラムの記述が容易になっているとともに、サンプリングのアルゴリズムも向上しています。

以下の文献では、MCMCプロシジャのRANDOMステートメントを含めたプログラムとともに、例題、詳細が記載されています。

The RANDOM Statement and More : Moving On with PROC MCMC

http://support.sas.com/resources/papers/proceedings11/334-2011.pdf

● NLINプロシジャ

非線形モデルの当てはめを診断するための統計量を出力するためのオプションNLINMEANSURES、BIASオプション(評価版)が追加されました。また、PLOTS=オプションを用いることによって、推定したモデル、適合度診断統計量などをODS統計グラフ機能にて描画できます。

● PHREGプロシジャ

新たにRANDOMステートメントが追加され、ランダム効果を含むモデル、例えばfrailty modelの分析に対応しています。また、NLOPTIONSステートメントが追加され、モデルの推定に用いる非線形最適化に関し、収束基準の設定などを指定することができます。

※ NLOPTIONSステートメントは、CALISプロシジャ、GLIMMIXプロシジャ、HPMIXEDプロシジャにてサポートされています。SAS 9.3ではPHREGプロシジャ、SURVEYPHREGプロシジャ、VARIOGRAMプロシジャでのサポートが追加されました。

● SURVEYPHREGプロシジャ

SAS® 9.22にて評価版として追加されましたが、SAS 9.3では正規版のプロシジャとして実行できます。

3.4 参考文献

以上ではSAS/STAT分析機能における主な拡張点をご紹介しました。より詳細な内容に関しては以下のドキュメントをご参照ください。

http://support.sas.com/documentation/cdl/en/whatsnew/64209/HTML/default/viewer.htm#statugwhatsnew93.htm

また、ご紹介しました分析機能以外における拡張点を確認されたい場合には、以下にてご参照いただけます。

http://support.sas.com/documentation/cdl/en/whatsnew/64209/HTML/default/viewer.htm#titlepage.htm

4 おわりに

今号では、SAS 9.3における出力に関し、デフォルトの設定における変更、SAS/STAT分析機能の主な拡張点に関してご紹介しました。特に新たに追加されているFMMプロシジャに関しては、具体的なプログラム、出力結果を含めた記述を行いました。分析をすすめる上で、新たな選択肢として、SAS 9.3における拡張点が参考となれば幸いです。

SAS Academic News

SAS アカデミック・ニュース

事例紹介、第6章では、統計手法の基礎であるt検定と分散分析についてご紹介いたします。
今号では、t検定及び分散分析の概要から、Enterprise Guideのオペレーション、
処理結果をもとにした解説と導き出される仮説や解釈について解説しております。
SAS四方山話では、実際に論文を執筆する際のパラグラフの作成方法や
論文全体のデザイン、論文発表までの全体の流れ等をご紹介しています。

事例紹介

コラム
「SAS四方山話」

事例紹介

高柳 良太
國學院大學 経済学部および人間総合科学大学 人間科学部 兼任講師

第６章 SAS® Enterprise Guide® を使ったｔ検定と分散分析

前回は「記述統計」メニューの、「分割表分析」を使ってクロス集計とχ二乗検定の説明をしました。前回のχ二乗検定は、虫歯の「ある」「なし」を学生の性別で比較して差があるかどうかを検討したものでした。

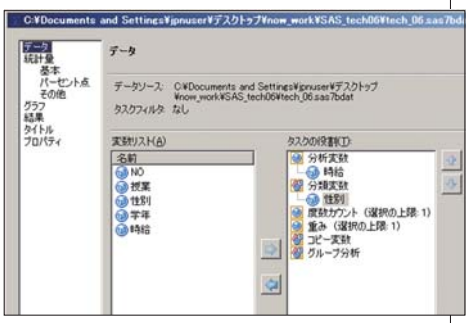
今回は、t検定と分散分析について説明をいたします。t検定と分散分析も、前回のχ二乗検定と同様に統計的検定であり、群（グループ）間で差があるかどうかを知りたいときに用いられます。χ二乗検定と違うのは、t検定や分散分析の場合は量的なデータについて、平均値などに差があるかどうかを知りたいときに使用する分析ということです。

Enterprise Guideを使ったｔ検定

まずは、Enterprise Guide（以下EG）を使ったt検定の説明をしたいと思います。今回のデータは、学生調査のデータで、アルバイト代について調査したものです。私が担当する授業に、1年生の必修である情報処理基礎という科目と、2年生以降の選択科目である情報処理演習というものがあります。授業ごとに履修している学生に、アルバイトの時給がいくらなのかを尋ねてみました（授業科目名やデータは架空のものです）。

まず、性別で時給が違うのかどうかを見てみたいと思います。雇用機会均等法もあり男女で時給の別があってはいけないのですが、ひょっとしたら違いがあるかもしれません。

単純に、性別ごとに時給の平均値を出したいのなら「記述統計量」の「要約統計量」で算出することができます。



上「記述統計」の「要約統計量」を選択
下「分析変数」に「時給」、「分類変数」に「性別」を指定

「記述統計」の「要約統計量」では、「分類変数」を指定すると、指定した変数ごとの統計量を算出します。

要約統計量
結果
MEANS プロシジャ

分析変数：時給

性別	オブザベーション数	平均	標準偏差	最小値	最大値	N
1	83	968.6746988	172.4484446	800.0000000	1500.00	83
2	77	952.5974026	162.2003520	800.0000000	1600.00	77

性別ごとの時給

これを見た限りでは、男子学生と女子学生は時給の平均は十数円しか違いません。この平均値の差について、統計的に

授業
1 情報処理基礎履修者
2 情報処理演習履修者

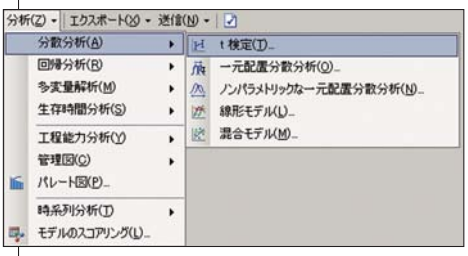
性別
1 男性
2 女性

学年
1～4 ただし情報処理基礎履修者は1年生のみ
情報処理演習は2～4年生

	NO	授業	性別	学年	時給
1	101	1	1	1	850
2	102	1	1	1	850
3	103	1	2	1	800
4	104	1	1	1	900
5	105	1			
6	106	1			
7	107	1			
8	108	1			
9	109	1			
10	110	1			
11	111	1			
12	112	1			
13	113	1			

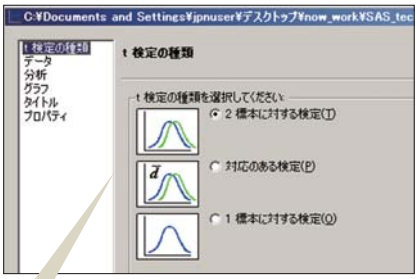
	NO	授業	性別	学年	時給
79	179	1	1	1	800
80	180	1	2	1	900
81	201	2	2	2	950
82	202	2	1	3	900
83	203	2	1	2	900
84	204	2	1	2	800
85	205	2	1	2	800
86	206	2	2	2	900
87	207	2	2	2	1000
88	208	2	2	2	900
89	209	2	2	3	1200
90	210	2	2	3	1250
91	211	2	2	3	900
92	212	2	1	3	800

有意差があるかをみたいと思います。このように2群（グループ）間に違いがあるかどうかを知りたい場合に、t検定を使用します。t検定は「分析」メニューの「分散分析」から「t検定」を選択します。



t検定

「分散分析」で「t検定」を選択すると、t検定の種類を選択する画面になります。

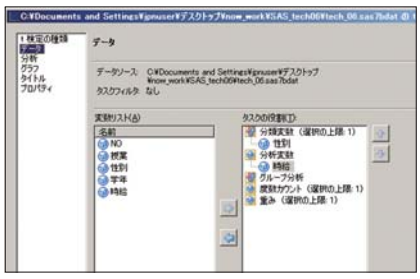


2標本に対する検定
2つのグループ間で平均値に違いがあるかというようなときに使います。

対応のある検定
例えば同じ人たちに2回調査を行って変化をみるなど、俗にペアのデータの場合の検定です。

1標本に対する検定
得られたデータが特定の集団から得られるデータと同じと判断できるかどうかを調べるものです。

今回は、男子学生と女子学生の違いなので、「2標本に対する検定」を選択します。



先ほどの記述統計量と同様に、「分類変数」に「性別」、「分析変数」に「時給」を指定して実行します。

t検定

TTEST プロシジャ

変数： 時給

性別	N	平均	標準偏差	標準誤差	最小値	最大値
1	83	968.7	172.4	18.9287	800.0	1500.0
2	77	952.6	162.2	18.4844	800.0	1600.0
Diff (1-2)		16.0773	167.6	26.5181		

性別	手法	平均	平均の95% 信頼限界		標準偏差	標準偏差の95% 信頼限界	
1		968.7	931.0	1006.3	172.4	149.6	203.6
2		952.6	915.8	989.4	162.2	140.0	192.8
Diff (1-2)	Pooled	16.0773	-36.2984	68.4530	167.6	151.0	188.4
Diff (1-2)	Satterthwaite	16.0773	-36.1777	68.3323			

手法	分散	自由度	t 値	Pr > |t|
Pooled	Equal	158	0.61	0.5452
Satterthwaite	Unequal	157.97	0.61	0.5443

等分散性				
手法	分子の自由度	分母の自由度	F 値	Pr > F
Folded F	82	76	1.13	0.5899

2標本のt検定は、分析している集団の分散が等しい場合と等しくない場合の2種類の分析があります。EGでは両方のt検定と、等分散性の検定結果を出力します。一番下に表示されている「等分散性」が等分散の検定の結果です。この結果から、t検定の結果のどちらを見るかを判断します。今回は「等分散性」の「Pr>F」が0.5899と5%（0.05）より大きいので、「分散が等しい」という帰無仮説が棄却されません。つまり、今回の男女の学生の時給データは分散が等しいと仮定できます。

等分散が仮定される場合は、上の出力で手法が「Pooled」のところを見ます。等分散が仮定できない場合は下の「Satterthwaite」を見ます。今回は「Pooled」の「Pr>|t|」が0.5452と5%（0.05）より大きいので、「男女で時給の平均が等しい」という帰無仮説が棄却されません。つまり今回のデータでは、性別では時給に違いがないことがわかりました。では、履修している授業の学生ごとで、時給が違うかどうかを見てみるとどうなるでしょうか。今度は「分類変数」を「授業」にして、「分析変数」は「時給」のままで実行します。

授業	N	平均	標準偏差	標準誤差	最小値	最大値
1	80	912.5	120.5	13.4747	800.0	1300.0
2	80	1009.4	192.4	21.5149	800.0	1600.0
Diff (1-2)		-96.8750	160.6	25.3862		

授業	手法	平均	平均の95% 信頼限界		標準偏差	標準偏差の95% 信頼限界	
1		912.5	885.7	939.3	120.5	104.3	142.8
2		1009.4	966.6	1052.2	192.4	166.5	227.9
Diff (1-2)	Pooled	-96.8750	-147.0	-46.7350	160.6	144.6	180.4
Diff (1-2)	Satterthwaite	-96.8750	-147.1	-46.6612			

手法	分散	自由度	t 値	Pr > |t|
Pooled	Equal	158	-3.82	0.0002
Satterthwaite	Unequal	132.71	-3.82	0.0002

等分散性				
手法	分子の自由度	分母の自由度	F 値	Pr > F
Folded F	79	79	2.55	<.0001

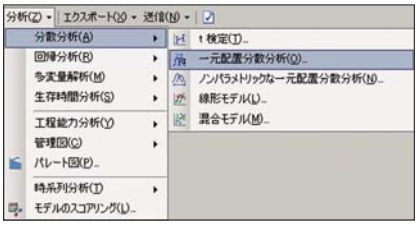
今回は「等分散性」の「Pr>F」が「<.0001」と非常に小さく、「分散が等しい」という帰無仮説が棄却されます。つまり、授業ごとの時給データは分散が等しいとは仮定できないことになります。等分散が仮定できないので、上の出力で手法が「Satterthwaite」のところを見ます。今回は「Satterthwaite」の「Pr>|t|」が0.0002と小さいので、基礎を履修している学生と演習を履修している学生では、時給の平均が等しくないということになります。その上の表に授業ごとの時給の平均値が算出されており、これを見ると基礎の学生よりも演習の学生の方が時給が高いことがわかります。

分散分析

t検定で、男女に時給の違いがなく、履修している授業での学生で時給に違いがあることがわかりました。ただ、履修している授業といっても、基礎の方は1年生の必修科目で、演習は2年生以上用の科目です。単純に考えても、2年生以上の学生は長い期間アルバイトをしている可能性が高いので、その分昇給している可能性があります。では、学年ごとに違いを見ていくことを考えてみましょう。

t検定は2つの集団の違いを見ることしかできません。3つ以上の集団の違いを知りたい場合は、分散分析を使用します。EGではt検定も分散分析メニューに入っており、t検定も分散分析も目的は似たような分析です。違うのは、t検定は平均値の差で比較をするのに対して、分散分析は全体の分散とグループごとの分散の比率の違いを見ています。平均値の差で比較する場合はグループごとに比較しなくてはいけないので、3群（グループ）以上になると比較対象が多くなりすぎてしまい、精度の問題など実用には不向きです。このため、3群以上の比較を行う場合は、一度に比較ができる分散分析を使用します。

分散分析を行うには、「分析」の「分散分析」の「一元配置分散分析」を選択します。



分散分析の場合は、t検定における「分類変数」が「独立変数」に、「分析変数」が「従属変数」にあたります。今回の分析では、学年ごとに時給が異なるかどうかを見るので、「独立変数」に「学年」を、「従属変数」に「時給」を指定します。

分散分析は、グループ全体とグループに分けたときに分散に違いがあるのかということを見る分析です。そのため、分散分析はどのグループ間に差があるかということ知ることはできません。どのグループ間に差があるかということを見る場合は「多重比較」というものを行います。EGでは「平均」の「多重比較」で指定します。

多重比較にはいろいろな分析方法があり、学会等でもいろいろ意見がでるものです。また、方法によって分散分析の結果と併用できるものと分散分析とは独立して行うべきとされるものなどがあり、学生や統計になれていない人に理解してもらうにはなかなか苦労するところです。

もちろん、それらをきちんと理解してから分析を行うべきであるというのが正しい意見だと思います。しかし、すべての人にそこまで求める必要性があるかということも疑問です。個人的には使い方を間違っていなければ、細かい部分の理解は後からでいいと思っています。そうでないと、入試で数学を選択していない経済学部生などに対して、データを扱う授業が成り立たなくなってしまいます。

ここでは比較的簡単に使えるBonferroniの多重比較を使います。ただ、Bonferroniは分散分析とは異なる検定なので、必ずしも分散分析の結果と一致しないこともあります。つまり、分散分析では有意差があるのに多重比較でBonferroniを使用したところ、有意差が見られる群の組み合わせがなかったということがあり得ます。また、Bonferroniで分析をするのであれば、あえて分散分析をする必要はないという話もあります。

ただ、EGをはじめ、ほとんどの統計パッケージの分散分析メニューに多重比較がオプションの形で存在しています。従って、多重比較のみを単独で行うことはあまりないというのが現状だと思います。ここでは、分散分析のオプションとして、Bonferroniの多重比較を実施する形をとりたいと思います。

要因	自由度	平方和	平均平方	F 値	Pr > F
Model	3	482645.230	160881.743	6.33	0.0004
Error	156	3965714.145	25421.245		
Corrected Total	159	4448359.375			

R2 乗変動係数	Root MSE	時給 の平均
0.108500	16.59217	159.4404	960.9375

要因	自由度	Anova 平方和	平均平方	F 値	Pr > F
学年	3	482645.2300	160881.7433	6.33	0.0004

一元配置分散分析を実行すると、上のような分散分析表が出力されます。一番上の分散分析表の1行目（Model行）と、一番下の表は同じです。一番右の「Pr>F」が有意確率です。

結果を見ると、「Pr>F」が0.0004と非常に小さい値なので学年間で時給に差がないという帰無仮説は棄却され、学年間で時給の平均値に違いがあると考えて良さそうです。

前述のように、どのグループ間に差があるかということを見る場合は「多重比較」の結果を見ます。

アルファ	0.05
誤差の自由度	156
誤差の平均平方	25421.24
t の棄却値	2.67232

有意水準 0.05 で有意に差があることを *** で示しています。

学年 比較	平均の差	同時 95% 信頼限界		
4 - 3	6.48	-185.82	198.78	
4 - 2	79.61	-105.10	264.32	
4 - 1	145.83	-34.52	326.18	
3 - 4	-6.48	-198.78	185.82	
3 - 2	73.13	-29.76	176.02	
3 - 1	139.35	44.52	234.18	***
2 - 4	-79.61	-264.32	105.10	
2 - 3	-73.13	-176.02	29.76	
2 - 1	66.22	-12.08	144.53	
1 - 4	-145.83	-326.18	34.52	
1 - 3	-139.35	-234.18	-44.52	***
1 - 2	-66.22	-144.53	12.08	

多重比較の表は、群ごとの組み合わせ間に有意差があるかどうかを調べます。多重比較によって組み合わせの設定が異なり、特定の群と他の群の組み合わせのみしかしないものなどがあります。Bonferroniの場合、すべての群の組み合わせを比較します。Bonferroniが比較的使いやすい多重比較と言われるのは、こうしたこともあります。ただし、群の数が多くなり組み合わせが増えると、Bonferroniは精度が落ちるとも言われているので注意した方がよいでしょう。

すべての組み合わせを比較するので、

この表の場合は上から順に「4年生と3年生の比較」「4年生と2年生の比較」となっており、「1年生と3年生の比較」「1年生と2年生の比較」で終わります。当然、「4年生と3年生の比較」と「3年生と4年生の比較」は同じものです。各数値はどちらの平均値からどちらを引くかで計算されるので、各値の絶対値が異なるだけです。

今回のデータでは、3年生と1年生に有意差ありとなりました（表では「3-1」と「1-3」）。つまり、1年生と3年生ではアルバイトの時給の平均値に違いがあったということになります。平均値の差は約139円です。

それ以外の学年間では、アルバイトの時給の平均値に差があるとはいえない結果となりました。1年生と2年生は約66円の差がありますが、それは統計的な有意差ではなく誤差の範囲内ということになります。同様に2年生と3年生も約73円の差がありますが、これも統計的な有意差とはいえないということです。

また1年生と4年生は約146円の差がありますが、有意差ありとはなりませんでした。1年生と3年生の平均値の差は、139円でした。それより大きな差でしたが、4年生は6人しかいなかったので測定誤差が大きく、有意差ありとはならなかったようです。4年生は他の学年に比べてあまりに人数が少ないので、はじめから分析から除外した方がよかったかもしれません。

一元配置分散分析は、このように3つ以上のグループ間で比較を行いときに使用しますが、グループの人数が極端に少ない場合や、グループの数そのものが多すぎる場合は使用に気をつけなくてはいけません。

次回は重回帰分析などについて説明をしたいと思います。

コラム「SAS四方山話」

大橋　渉
千葉大学医学部疾患プロテオミクス寄附研究部門　データベース・インフォマティクス担当
ヤンセンファーマ株式会社　研究開発本部　臨床統計部　統計解析グループ　マネージャー　医学博士

第15回 医学研究のススメ（3）

いきなり書けない…方々のために

前回は、研究計画の立案や心構えなどについて書かせていただきましたが、本シリーズを読まれた方より、「日本語論文も書けないのに、どうしていきなり英文論文なんて書けるんですか？」との手厳しいご意見を頂きました。確かにこの方のおっしゃることもよく分かりますが、筆者の知る中には、初めて英語論文を書いて掲載にこぎつけた人も多数います[*1]し、自然科学・社会科学分野の学術雑誌を対象として、その雑誌の影響度を測る指標であるImpact Factor（以下IF）が10点台の雑誌に掲載などという、非常に羨ましい人もいます。そういう意味でも、筆者としては是非とも多くの方々に論文発表にチャレンジしていただければな、と考えている次第です。

英語が苦手なら手始めに日本語で、と思ったとしても、IMRAD構造などは全て英語論文と共通ですので、執筆の難易度や困難さが急激に低下するわけではありません。日本語で執筆したものをプロの翻訳家に任せるなどの方法もございますので、やはり言語よりも内容が物を言います。もっとも、参考文献として大量の英文論文を読み込まなければならないことだけは間違いありませんので、研究活動にはそれなりの英語力が必要であることは確かです。

1）まずはどんな形でも発表を

いきなりの論文化が厳しいという方は、まずはご自身のご興味や関心などの発表を、所属機関の中で行ってみてはいかがでしょうか？ 研究論文などという大げさなものではなくとも、ご自身の経験や体験（症例報告：Case Reportのようなもの）、気付きなどを、可能な限りIMRAD構造に沿う形で記載してみましょう。それをまとめて要点をスライドで作成し、15～20分程度の発表を行ってみるのも良いかもしれません。ご自身のテーマに対するCommunityの中での評価も知ることができますので、一人で考えていても前に進まなかったことが、何か見えてくるかもしれません。

2）（Next Step）　学会発表の流れ

皆様がこの記事をお読みになっている頃は、まさしく「学会シーズン」であり、毎週どこかで何らかの学会が開催されているでしょう。実は一度学会発表をすることで、研究論文の投稿と同様のStep（詳細は次号にて）を体験することができますので、論文投稿前に一度はデビューしてみるのも良いと思います[*2]。

よし、ならば早速今年、とは残念ながらいきません。なぜなら、多くの学会は秋の発表者を春先に募集し、半年かけて査読を行うことで質の確保や発表者の絞り込みを行うからです。応募者はほぼ全員発表できる（ポスター含め）学会もあれば、**論文以上に査読が厳しい学会もあります**ので、時にRejectの憂き目に遭ってしまうこともあります。しかも論文のように「別の雑誌に」というわけにもいかない場合（時期的に）も多いので、来年か、早くて半年後か、下手をすればお蔵入りなどということもないわけではありません。参考までに筆者の場合ですが、何本かお蔵入りになっております[*3]。次項の図1は、簡単な学会発表までのフローです。

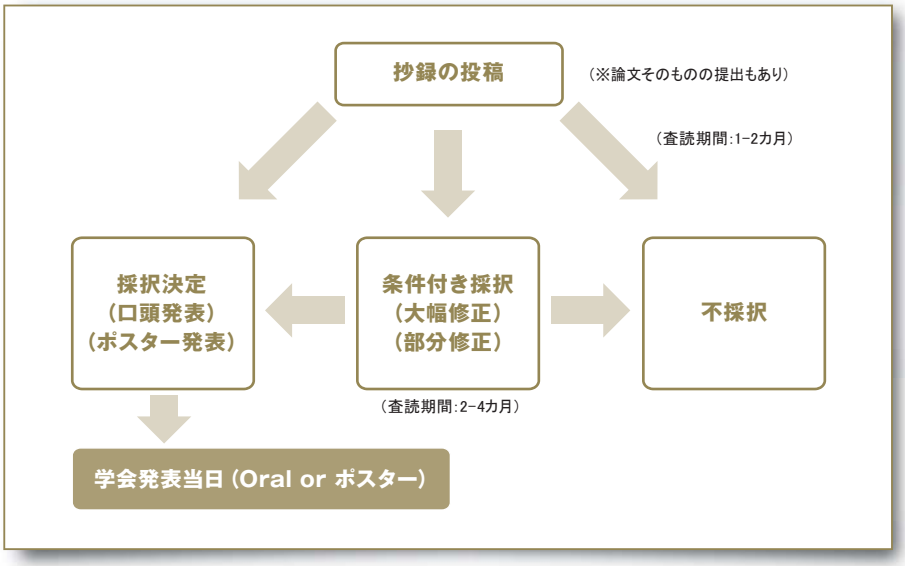


図1 学会発表までの流れ

3）論文の顔 Abstract

まずは最初の抄録ですが、最近はほとんどがオンラインによる投稿です。通常は数百文字程度で、論文で言うところの**Abstract（概要）**に近い記載をすれば良いのですが、時に論文そのものの形で出すように求めてくる学会もあります。なお、Abstractは、Pubmedなどの論文検索サイトにおいて、どのような論文でもここだけは**無料で読める部分**ですので、自身の論文をたくさんの人に読んでいただくためには、実は非常に重要な部分です。多くの皆様は、Abstractを読んで論文の本編を読むか読まないかを決定しているので、是非とも目一杯「見せるための」工夫をしてください。また、論文の査読者もここを念入りに読みますので、まさしく**研究・論文にとっては顔ともいえる部分**です。

研究開始前に抄録を記載して、学会当日までにそれに合わせて研究を進めるというパターンもありますが、筆者の経験則ではやはり研究終了後に抄録を記載した方が上手く書けます。なお、オンラインですので締め切り間際になると大変混雑し、つながり難くなりますので、余裕を持って投稿しましょう。

4）Oral（口頭発表）かポスターか？

次に発表の種類ですが、Oral（口頭発表）とポスターの2種類があります。Oralの場合はMicrosoft Power Pointなどでスライドを作成して、20 ～ 30分程度の発表と10分程度の質疑応答を行います。ポスターの場合ですが、既定の大きさの用紙（基本的にはA0です）に研究に関する全ての情報を記載し、原則として掲載時間内はポスターで待機して、閲覧者の質問に回答しなければなりません。時間的制約やインパクトの観点からも、どの学会も基本的にはOralの希望者が多くなる傾向がありますので、Oral発表がRejectでポスター発表になってしまうことは、かなりの頻度で発生します。いきなりポスターもRejectというパターンはあまり聞いたことがありませんが、Oral条件付き再審査、最終的にRejectといったパターンは頻繁にあります。「**あくまでOralにこだわるか、ポスターでも良いか**」のような“駆け引き”も、時には必要になってきます。

再審査に関しましては、査読者の疑問に対し「これでもか！」というぐらい徹底的に返答しましょう。時に査読者が間違っていることもあり得ますので、そんな時には大いに反論してください[*4]。近い将来、それらの経験は論文の査読者とのやりとりで必ず役に立ちます！

Method（研究方法）の記載を読む

さて、前回までにIMRADの各項目の注意点について記載させていただきましたが、中でも特に注意すべき点が多い項目としてM（Method：方法）を挙げさせていただきました。もしも統計処理を行っているのであればその解析方法を、割り付けを行っているのであればその方法を、必要症例数の見積もりを行っているのであれば、その根拠となる数値を記載しなければなりません。ここからは実際の論文を事例として進めていきましょう。

以後、出典は全て「Ann Rheum Dis 2009; 68:789–796. doi:10.1136/ard.2008.099010. **Golimumab, a human antibody to tumour necrosis factor a given by monthly subcutaneous injections, in active rheumatoid arthritis despite methotrexate therapy: the GO-FORWARD Study**」です。こちらの試験は、関節リウマチの薬物療法について、生物学的製剤であるgolimmab対プラセボ、対メトトレキサート（MTX）の比較を行っています。

PATIENTS AND METHODS

This was a phase **III, multicentre, randomised, double-blind, placebo** controlled trial. The study included a double-blind controlled phase to week 52 and an open-label extension up to 5 years. In this report, **we present the results to week 24**, which include the co-primary endpoints at weeks 14 and 24. **Patients were enrolled at 60 investigational sites in 12 countries:**

患者背景と試験方法の要約

・第3相試験、多施設、無作為割り付け、二重盲検対プラセボ試験（52週まで）
・結果は24週時点のもの
・12カ国、60施設で実施

こちらの論文の場合は、IMRADのM（Method：方法）が「Patients and Methods」となっておりますが、この場合は患者背景の説明を含んでいます。ここでは省略しておりますが、「18歳以上で登録の3カ月前から関節リウマチを罹患していて…」などの条件が記載されています。

必要症例数に関する考察

前々回の連載にて、検証型の臨床試験を行うためには「症例数の設定」が必要であることを記載させていただきました。これは、①必要以上に患者さんを危険な目に遭

わせない、②限られた時間と予算の範囲で研究を行わなければならない、③意味の変化や差に反応させない（統計的有意＝臨床的有意ではない）ー以上の3点が理由でした。というわけで、科学的批判に耐えられる（＝臨床試験の結果を担保できる）だけの、必要最低限の症例数を計算しなければなりません。必要症例数を設定するためには、「**主要評価項目**」と「**設定根拠（＝見積もられる臨床的差異）**」が必要でしたが、まずはこちらの研究における主要評価項目から確認してみましょう。

Evaluations

Response to treatment was assessed using the ACR response criteria (ACR20/50/70).[19] ACR-N[9 20] was also calculated.

評価項目

・ACR20、ACR50、ACR70を用いる。
・American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trialsの略。たとえばACR20とは、20%以上の改善が見られた患者の割合を示す。

どうやら主要評価項目は、関節リウマチの改善度の指標であるACR20です。ということは、評価項目はどうやら「割合の差」になりますので、**χ二乗検定**による比較を行わなければなりません。χ二乗検定に必要な症例数の設定は、以下の公式によって算出することが可能です。

$$n=\frac{\left\{Z_\alpha\sqrt{2\bar{P}(1-\bar{P})}+Z_\beta\sqrt{P_1(1-P_1)+P_2(1-P_2)}\right\}^2}{(P_1-P_2)^2}\left(\bar{P}=\frac{P_1+P_2}{2}\right)$$

（n：必要サンプル数、P1：被験薬群の反応率、P2：対照群の反応率、Z_α：有意水準から算出する値（通常5%で1.96）、Z_β：検出力から算出する値（通常80%で0.84））

では、続いて「設定根拠」に注目してみましょう。こちらはタイトルもそのままで「症例数設定」とございますので、非常に分かりやすいです。

Sample size calculation

Assuming 55 % or more of patients in groups 3 and 4 and 35 % of patients in group 1 would achieve an ACR20 response, a sample size of 120 patients in group 1 and 80 patients in groups 3 and 4 was required to achieve greater than 90 % power (two-sided x_2, a = 0.05). Assuming 55 % of patients in group 2 and 35 % of patients in group 1 would achieve an ACR20 response, a sample size of 120 patients in both groups 1 and 2 was needed to achieve greater than 85 % power (two-sided x_2 test, a = 0.05).

症例数設定根拠

1.ACR20に該当する割合が、群3及び群4において55%以上、群1において35%であると仮定すれば、群1（n=120）、群3（n=80）、群4（n=80）における検出力は90%以上（両側検定、有意水準α=0.05の場合）となる
2.同様に群2で55%、群1で35%の場合、両群ともn=120であれば検出力は85%以上（両側検定、有意水準α=0.05の場合）となる

しかしながらこちらを読んでみますと、既に症例数は決まっているようです。見込まれる差から必要症例数を求めるというような通常のスタイルではありませんが、ここでは研究結果の担保が可能であるかを確認するために検出力を求めるパターンです。既に群1（n=120）、群2（n=120）、群3（n=80）、群4（n=80）と症例数が決まっており、さらにそれぞれ反応率の差は少なくとも55-35=20%と設定されています。

これらの前提に基づいて、それぞれの検出力（Power）を求めてみましょう。

プログラム（1）

```
proc power;
  twosamplefreq test=pchi
  grouppproportions = (.55 .35)      /*群3、4の反応率は最低でも55%、群1は35%*/
  groupns = 80 | 120                 /*群3、4は80症例、群1は120症例*/
  power =.;
run;
```

```
The POWER Procedure
Pearson Chi-square Test for Two Proportions

Fixed Scenario Elements

Distribution                    Asymptotic normal
Method                          Normal approximation
Group 1 Proportion                             0.55
Group 2 Proportion                             0.35
Group 1 Sample Size                              80
Group 2 Sample Size                             120
Number of Sides                                   2
Null Proportion Difference                        0
Alpha                                          0.05

Computed Power

    Power
    0.802
```

群3または4の反応率
群1の反応率
群3または4の症例数
群1の症例数
両側
有意水準（α）
検出力：80.2%

図2 プログラム(1)の実行結果

この場合の検出力は 80.2%となったわけですが、こちらの記載によれば群 3、4 における反応率は 55%以上と期待されており、その場合に 90%以上の検出力としているわけです。80%という数値自体も非常に説得力があり、通常用いるには十分な検出力ですので特に筆者としても意義はございませんが、こちらで主張される 90%以上には若干不足しているようです。ならば、群 3、4 の反応率が何%以上であれば、検出力は 90%以上を確保できるのでしょうか。少し確認してみたいと思います。

プログラム (2)

```
proc power;
  twosamplefreq test=pchi
  proportiondiff = 0.20,0.25,0.30,0.35          /* 群 1 の 35%との反応率の差 */
  refproportion = 0.35                          /* 群 1 の反応率 35%と上記の全てを比較 */
  groupns = 120 | 80
  power = .;
run;
```

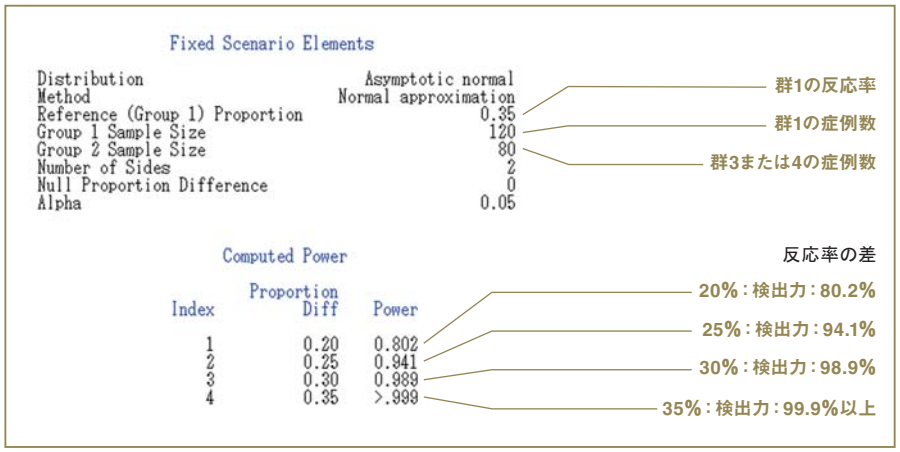


図3 プログラム(2)の実行結果

群3または群4の反応率が35%+20%=55%では、プログラム(1)の結果同様、検出力は80%程度となります。症例数設定根拠の1.では、「これが55%以上のときに90%以上の検出力となる」としておりますが、実際に反応率が60%ならば検出力が94%になりますので、どうやらこの論文における主張は間違っていないようです。

最後に症例数設定根拠の2.を見てみましょう。反応率の差は35%対55%と1.と同様なのですが、症例数は両群とも120症例です。

プログラム (3)

```
proc power;
  twosamplefreq test=pchi
  groupproportions = (.55 .35)
  npergroup = 120                               /* 両方とも 120 症例 */
  power = .;
run;
```

プログラム(3)の出力項目は、基本的にプログラム(1)と同じですので結果の解説は省略致しますが、検出力は88.1%となります。1.の場合と同じ反応率の差にも関わらず、8%程度ですが検出力は増加しております。これはもちろん、片方の群において症例数が40症例多いということが原因です。また、他の条件が全て同一であれば、症例数は同じ場合の方が検出力は高くなります。

次回、このシリーズの最終回を予定しております。IMRADの他の部分や参考文献、さらには査読者とのやりとりに関するお話などをさせていただく予定です。

*1) 誤解のないように申し上げますが、英語論文が日本語論文よりも優れているとか、掲載が難しいなどと言っているのではありません。IFは付与されていなくても、日本語で優れた論文もございますし、一方では「ただ英語だけ？」のような論文もあることはあります。時には、大勢の人の目につくように「戦略的に」あえて日本語論文に投稿する研究者もいます

*2) ただし、**ネタに関しましては、（もしも論文化する予定があるのであればですが）論文化前の、機密性の高いネタは避けましょう。**競合する研究者に真似をされてしまったというお話も耳にします

*3) などと言いつつも、教育系の学会で10年前のネタを発表したこともあります。自然科学系では10年前のネタは確実に**時代遅れ**と見なされてしまいますが、人文・社会科学における理論などでは必ずしも時代遅れとは見なされない、不易と流行という意味での「不易」の部分に対するニーズはあります

*4) しかし、以前筆者がある雑誌の査読者に対し意義を唱えたところ、Rejectされてしまいました（内容：A-C群の3群の比較において多重比較による調整を行ったところ、「t検定を3回繰り返すように」と指示があったため、その必要性がないと訴えた）

Q&A

Q 複数変数に文字列が格納されています。特定の文字列が含まれている変数を特定したいのですが、何かよい方法はありませんか。

A SAS® 9.2の新機能であるWHICHC関数により、変数間の文字列検索が容易に行えます。

例

```
DATA _NULL_;
  var1=" 林檎 ";
  var2=" 梨 ";
  var3=" メロン ";
  x1=WHICHC(' 林檎 ',of var1-var3);  /* WHICHC 関数 */
  x2=WHICHC(' メロン ',of var1-var3);
  x3=WHICHC(' 洋梨 ',of var1-var3);
  PUT x1=;
  PUT x2=;
  PUT x3=;
RUN;
```

ログの出力結果

```
-------------------------
  x1=1
  x2=3
  x3=0
-------------------------
```

Q 64ビット版SAS 9.3にて以下のプログラムを使用してExcelファイルをインポートすると、エラーが表示されます。

```
PROC IMPORT DATAFILE="c:¥temp¥tst_space1.xls"
OUT=out DBMS=EXCEL REPLACE;
RUN;
```

```
ERROR: DBMS タイプ EXCEL は、import には無効です。
```

これはなぜでしょうか。原因について教えてください。

A 64ビット版SAS 9.3のExcelエンジンは64ビット版Excelが必須になります。そのため64ビット版SAS 9.3と32ビット版のExcelがインストールされている環境にて、Excelファイルのインポートを行った場合は上記エラーが表示されます。該当する場合は64ビット版のExcelをインストールするか、SAS PC Files Serverをご利用ください。SAS PC Files Serverにつきましては、SAS 9.3 DVD-ROMよりインストールすることが可能です。

お使いのExcelのビットに関してはExcelのオプションよりご確認いただけます。
Excel 2007をお使いの場合:
Officeアイコンをクリック-「Excelのオプション」-「リソース」-「バージョン情報」
Excel 2010をお使いの場合:
「ファイル」-「オプション」-「言語」

Q DI Studioで列のプロパティで出力形式のメニューをプルダウンしても、ユーザ定義フォーマットが選択肢として表示されませんが、使用できないのでしょうか。

A DI Studio で既存データセットの出力形式のプルダウンメニューにユーザ定義フォーマットが選択肢として表示されるのは、
・もともとデータセットに該当のフォーマットが適用された状態でテーブルメタデータを作成していた場合
または
・ユーザが手入力でフォーマット名を出力形式欄に記入して保存した場合
のいずれかになります。
従って、出力形式が定義されていない、または標準のフォーマットが適用済みである変数の出力形式欄をプルダウンしても、ユーザ定義フォーマットは選択肢に表示されません。これは仕様になります。
上記のような変数についてユーザ定義フォーマットを適用したい場合は、手入力にてフォーマット名を入力していただく必要があります。

Q ライブラリを定義する際に、事前にライブラリフォルダを準備する必要があります。そのフォルダがない場合、OS上でそのフォルダを新規作成することなく、SASからそのフォルダを作成できませんか。

A SAS® 9.3より、ライブラリ定義をする際に指定したライブラリフォルダが存在しない場合、そのフォルダを作成するDLCREATEDIRシステムオプションが新たに追加されました。

例

```
OPTIONS DLCREATEDIR ;
LIBNAME sample  "c:¥new";
```

ただし、作成先のフォルダには実行ユーザーの書き込み権限が必要となります。

詳細に関しましては、次のWebページに記載されております。
http://support.sas.com/documentation/cdl/en/lesysoptsref/63325/HTML/default/n1pihdnfpj4b32n1t62lx0zdsmdn.htm

Q SAS® 9.3から、POWERプロシジャ、GLMPOWERプロシジャにてもODS統計グラフ機能がサポートされています。しかしながら、複数の条件に対するグラフでは、シンボル、線種、色が同じとなります。設定した各条件との対応がわかりやすいよう、異なるシンボル、線種、色に変更できますか。

A ODS統計グラフ機能を用いている場合、ODSスタイルにてシンボル、線種、色などが設定されます。デフォルトのODSスタイルはHTMLBLUEとなりますが、POWERプロシジャ、GLMPOWERプロシジャのグラフ表示には適していません。
このため、これらのプロシジャにてODS統計グラフ機能を用いる場合には、以下のステートメントを用い、明示的に異なるODSスタイル形式を指定します。

```
ODS HTML STYLE=htmlbluecml;
```

HTMLBLUECMLの他、適しているODSスタイルはSTATISTICAL、ANALYSIS、DEFAULT、 LISTINGとなります。

Q SAS® に適用済みの Hot Fix の一覧を確認する方法はありますか。

A **SAS V8、SAS® 9.1.3 の場合**

1. 下記ページの "Download" タブから、
SAS InstallReporter3.sas をダウンロードします。
http://support.sas.com/kb/20/390.html

2. SAS を起動し、SASInstallReporter3.sas を実行します。

3. アウトプットウィンドウに結果が表示されます。
「3: Status of Hot Fixes found for SAS X.X」配下に適用済み Hot Fix 一覧が表示されます。

SAS® 9.2 、SAS® 9.3 の場合

1. <SASホームディレクトリ>¥deploymntreg¥ 配下に
sas.tools.viewregistry.jar が存在する事を確認します。

例

```
C:¥ProgramFiles¥SAS¥deploymntreg¥sas.tools.viewregistry.jar
```

2. 存在しない場合、下記ページの "Download" タブから、
sas.tools.viewregistry.jar をダウンロードします。
http://support.sas.com/kb/35/968.html

ダウンロードした sas.tools.viewregistry.jar を以下のフォルダ直下に保存します。

```
<SAS ホームディレクトリ>¥deploymntreg¥
```

例

```
C:¥Program Files¥SAS¥deploymntreg
```

3. sas.tools.viewregistry.jar を実行します。Windows の場合はダブルクリックで実行可能です。

4. 同フォルダ内に DeploymentRegistry.html が作成されます。
ブラウザで開き、"Hot Fix Entry" で検索いただくことで適用済み Hot Fix を確認できます。

SAS Training

SAS トレーニングのお知らせ

祝! SAS® グローバル認定プログラム 日本語化 1周年

2010年9月1日にSASのプログラミング系認定試験である、SAS® Base Programming for SAS® 9(以下:Base試験)とSAS® Advanced Programming for SAS® 9(以下:Advanced試験)が日本語で受験できるようになり、早1年が経ちました。その後、SASプラットフォーム製品に対応する3資格も追加され、ほとんどのSAS認定資格は日本語で受験できるようになりました。この1年で、日本のSAS資格者数は倍増し、金融・医薬・製造・通信・サービスなど幅広い業種の中で、SAS認定プロフェッショナルが活躍しています!

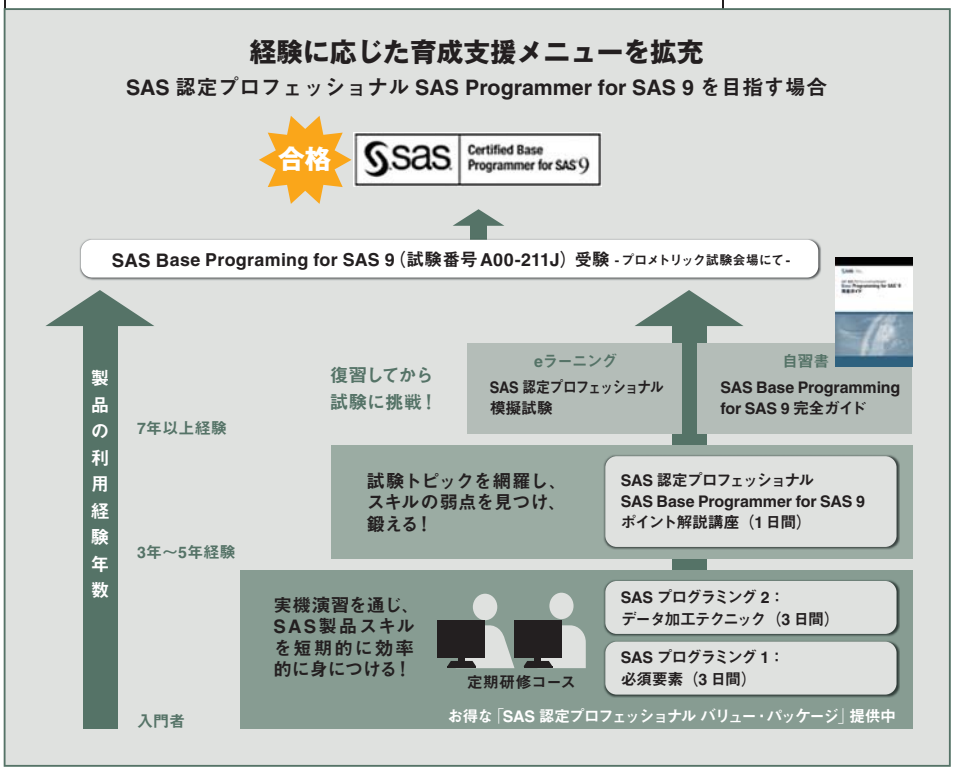
日々の業務の中では、既存のSASプログラムを編集し、使い慣れたプロシジャやテクニックに偏りがちになります。ある合格者の方が、「合格はもちろん嬉しかったけれど、受験勉強する道のりの中で身についたバランスの良い知識とテクニックが、一番の収穫だったと思う。」とお話しされていました。私どもも、資格取得はゴールではなく、スキルアップの手段の一つとして上手に活用してもらいたいと思います。最近は、SAS製品への対応力を客観的に計る指標の一つとして、SAS認定プロフェッショナルをご要望される企業様の声も聞こえるようになり、嬉しく感じています。

SAS経験に応じた学習メニューを拡充

「深いナレッジを持つ日本のSAS利用者の皆さまに、世界に通用するSAS資格も是非とってほしい!」と願い、資格取得を支援するメニューの拡充に務めた1年でした。
SAS資格を支える礎は、Base試験です。試験に挑戦する80%以上の方々が、SAS Foundationの普遍的な技術要素を問う「Base試験」を、初めてのSAS受験科目に選ばれます。
これからSASプログラミングを習い始める方もすでに十分な経験をお持ちの方も万全な備えで試験に臨めるように、定期トレーニングだけでなくポイント解説講座や自習書eラーニングの模擬問題など、さまざまな学習メニューを揃えました。

Base資格保持者のネクスト・キャリアパス - その1

Base試験に合格された方にお勧めするのは、Advanced試験です。Advanced試験では、SASマクロ機能、SAS SQLコードや高度な効率化テクニックなど、大規模なデータ環境で対応できるSASプログラミング力を問うとても実務的な間口の広い試験です。
難易度の高いAdvanced試験に挑戦をしようと考えている方に朗報です。
ご要望の声が高かったAdvanced試験のポイント解説講座を12月からスタートします。
Advancedポイント講座URL:
www.sas.com/jp/training/course/cert_advanced_seminar.html



SAS® 認定 プロフェッショナル SAS Advanced Programmer for SAS® 9 ポイント解説講座

初回12月26日開催分は、特別価格71,400円(本体価格68,000円)でご受講いただけます。

英語書籍「SAS Certification Prep Guide: Advanced Programming Programming for SAS 9, Third Edition」の練習問題を使用して、SAS Advanced Programming for SAS 9で必要な大部分のトピックの復習を行います。また、本コースには、受験チケット(バウチャー;18,900円(税込)相当)と英語版書籍($129相当)の両方が含まれています。

Base資格保持者のネクスト・キャリアパス – その2

医薬業界に従事される方には、Advanced試験だけでなく、もう一つのお勧めの資格が登場しました。2011年夏に米国で開始した新資格「SAS Clinical Trials Programmer Using SAS® 9（以下、CT試験）」が、2011年12月には日本語でも受験可能になります。
業界標準とも言えるSASによる臨床試験解析。治験データを、SASプログラムを用いて加工、レポート、そして解析処理へ。分析までのスキルを問うSAS資格が、ついに医薬分野で初登場です。Base資格、Advanced資格そしてCT資格、この3資格を兼ね備えたSAS認定プロフェッショナルは、医薬業界のSASエキスパートと言っても過言ではありません。

ここからは、CT試験の日本語化に備え、新試験の概要と、一足先にサンプル問題もご紹介します。
CT試験URL:
www.sas.com/jp/training/certify/ctp9.html

CT試験の概要

今回日本語で受験できるようになったCT試験は SAS Base Programming試験の合格者を対象とした試験です。試験時間は120分で、71問の出題中70%の正答率で合格となります。

試験範囲は表1に示すように、9つのカテゴリに分かれています（サブカテゴリなどの詳細は弊社ホームページ: www.sas.com/jp/training/certify でご確認ください）。
初めの2カテゴリでは統計解析計画書の解釈やCDISC標準などの業界知識が問われます。それ以降のSASプログラミングに関連した問題でも臨床プロセスに従ったデータ変換や検証に関する問題や臨床データを使用した問題が多く出題されるため、CT試験の合格はより実践的な能力を示す証となります。

カテゴリ	カテゴリ名	サブカテゴリの数
1	臨床試験プロセス	4
2	臨床試験データの構造	5
3	臨床試験データのインポートとエクスポート	1
4	臨床試験データの管理	2
5	臨床試験データの変換	5
6	臨床試験への統計プロシジャの適用	4
7	臨床試験のためのマクロプログラミング	3
8	臨床試験結果のレポート	2
9	臨床試験データのレポートのバリデート	4

表1: CT試験の試験範囲

試験範囲に該当するトレーニングは、SASプログラミング1、SASプログラミング2、SASによる統計解析、SASによる回帰分析、SASマクロ言語1、SASレポートと非常に多岐にわたりますが、それだけ取得し甲斐のある資格と言えます。ぜひとも資格取得にチャレンジしてください。
今回はレポートに関連するサンプル問題を掲載します。

練習問題

以下のSASプログラムをサブミットします。

```
proc format ;
  value dayfmt  1='Sunday'
                2='Monday'
                3='Tuesday'
                4='Wednesday'
                5='Thursday'
                6='Friday'
                7='Saturday' ;
run ;

proc report data=diary ;
    column subject day var1 var2 ;
    ＜コードをここに記述＞
run ;
```

DIARYデータセットでは、変数DAYに出力形式DAYFMTを適用します。変数DAYの出力形式を適用しない値の順番で表示するには、プログラム中の＜コードをここに記述＞にどのステートメントを記述しますか?

A. `define day / order 'Day' ;`
B. `define day / order order=data 'Day' ;`
C. `define day / order noprint 'Day' ;`
D. `define day / order order=internal 'Day' ;`

解説

レポートに出力する行の順番を変更するには、DEFINEステートメントのORDER=オプションを使用します。指定できるオプションは表2で、PROC REPORTのデフォルトはFORMATTEDです。この問題は出力形式を適用していない値の順に表示させたいため、ORDER=INTERNALを使用している選択肢Dが正解です。

オプション	説明
DATA	入力データセットに登場する順番
FORMATTED	出力形式適用値の順番
FREQ	度数値の昇順
INTERNAL	出力形式を適用していない値の順

表2: DEFINEステートメントのORDER=オプション

正解：D

新資格リリース記念　キャンペーン情報

新資格のリリースを記念して、お得なディスカウント制度や受験チケットのセット販売（期間限定）を開始いたしました。詳細は下記をご参照ください。

ディスカウント制度：
www.sas.com/jp/training/certset.html

SASグローバル認定プログラム：
www.sas.com/jp/training/certify/index.html

受験チケットお申し込み：
www.sas.com/jp/training/certify/order.html

特別トレーニング・コースの開催のご案内

●「製薬企業におけるSASプログラミング【SAS Programming in the Pharmaceutical Industry】」コース（1日間）

［日　　程］
2012年1月27日（金）10:00 ～ 17:00（東京会場）
［価　　格］
73,500円（税込）／※チケットのお取り扱いはありません。
［受講対象］
SASシステムによる臨床試験の統計解析またはDM業務に従事している方
臨床開発経験1年未満のSASプログラマの方から、SASプログラマを管理する方まで、幅広いレベルの方々を対象としています。
［コンテンツ］
Chapter 1: 環境
Chapter 2: 分類変数の作成
Chapter 3: データの読み込み
Chapter 4: 解析用データの作成と変換
Chapter 5: 表やリストの作成
Chapter 6: グラフの作成
Chapter 7: よくある解析の実行と結果の取得
Chapter 8: データの出力
Chapter 9: 臨床試験におけるSASプログラミングの将来
Chapter 10: リソース
［テキスト］
本コースは、下記英語版の書籍を使用し、日本語で説明を行います。なお、補助資料等はございません。テキストは当日弊社受付にてお渡し致します。
『SAS Programming in the Pharmaceutical Industry』
(SAS Press)
［Web Page］
www.sas.com/jp/training/course/pharm_prog.html

ディスカウント制度のご案内

【特別企画 半日ごとに選べる講座で、賢く学ぶ冬のスキルアップキャンペーン】
スキルアップはしたいけれど、業務が多忙で連続した日程のトレーニングコースへの参加は難しい、トレーニング予算の確保が難しい、そんな方々へ朗報です！　拡張目覚ましいグラフ機能。関心度の高いトピック別に、なんと3時間で簡潔に学べるコースを、各回特別価格26,250円（税込）にて準備いたしました！
全2～3回から構成されていますが、各回のみの受講も可能です。また、セットでのお申し込みはさらにお得です！！

●SASマクロマスター（全2回 / 16:00 ～ 19:00）

［日　　程］
第1回　11月16日（水）
「マクロ・アプリケーションとは　～ SASプログラムの汎用化～」
第2回　12月 2日（金）「動的なアプリケーション」
※本セットは定期コース「SASマクロ1：必須要素」（115,500円（税込）/ 2日間）からの抜粋版となります。網羅的にじっくり学びたい方は、定期トレーニングの受講をお勧めいたします。
［価　　格］
セット価格　42,000円（税込）//　各回価格　26,250円（税込）

●SAS上級プログラミング・マスター（全3回 / 16:00 ～ 19:00）

［日　　程］
第1回　12月 8日（木）「I/O処理の効率化」
第2回　12月13日（火）「インメモリ結合による効率化」
第3回　12月22日（木）「大容量データの結合テクニック」
※本セットは定期コース「SASプログラミング3：上級テクニックと効率化」（173,250円（税込）/ 3日間）からの抜粋版となります。網羅的にじっくり学びたい方は、定期トレーニングの受講をお勧めいたします。
［価　　格］
セット価格　63,000円（税込）//　各回価格　26,250円（税込）
［Web Page］
www.sas.com/jp/training/skillup_winter11.html

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。
www.sas.com/jp/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: 03-6434-3690　　F A X: 03-6434-3691
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 / 9.2 / 9.3
64-bit Windows(Itanium)版	SAS 9.1.3 / 9.2 / 9.3

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 / 9.2/ 9.3

UNIXプラットフォーム

SunOS/Solaris版	SAS 9.1.3 / 9.2 / 9.3
HP-UX版	SAS 9.1.3 / 9.2 / 9.3
HP-UX(Itanium)版	SAS 9.1.3 / 9.2 / 9.3
AIX版	SAS 9.1.3 / 9.2 / 9.3
Linux(Intel) 版	SAS 9.1.3 / 9.2 / 9.3

SAS Technical News入手

SAS Technical Newsは、右記のURLから入手できます。 http://www.sas.com/jp/periodicals/technews/index.html

発行:SAS Institute Japan株式会社

STN SAS Technical News AUTUMN 2011

■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:03-6434-3680 FAX:03-6434-3681

SAS Institute Japan株式会社

本社
〒106-6111
東京都港区六本木6-10-1
六本木ヒルズ森タワー 11F
Tel 03(6434)3000
Fax 03(6434)3001

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

www.sas.com/jp