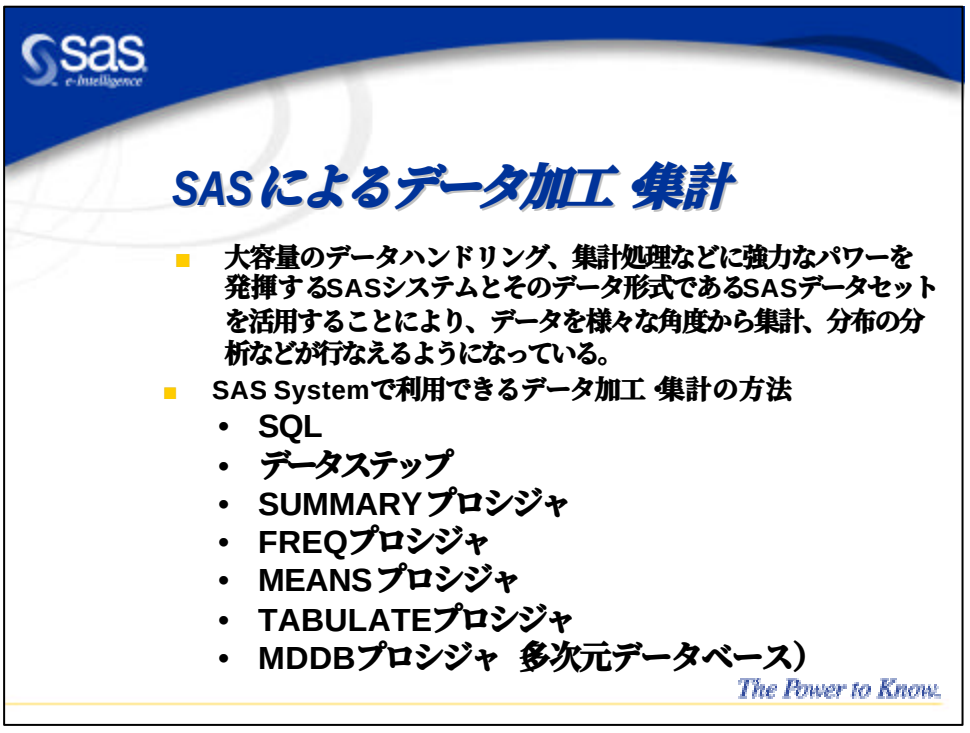

sas
e-Intelligence
「データマイニングの前の
proc freq とproc means」
SAS Institute Japan Ltd.
営業本部ソリューションプランニングセンター
Kazunari Azuma 東 一成）
jpnkaa@jpn.sas.com
The Power to Know.
Copyright © 2000, SAS Institute Inc. All rights reserved.


sas
e-Intelligence
SASによるデータ加工 集計
大容量のデータハンドリング、集計処理などに強力なパワーを発揮するSASシステムとそのデータ形式であるSASデータセットを活用することにより、データを様々な角度から集計、分布の分析などが行なえるようになっている。
SAS Systemで利用できるデータ加工 集計の方法
SQL
データステップ
SUMMARYプロシジャ
FREQプロシジャ
MEANSプロシジャ
TABULATEプロシジャ
MDDBプロシジャ 多次元データベース）
The Power to Know.

# MEANSプロシジャとFREQプロシジャ

FREQプロシジャ概要

- FREQプロシジャでは1次元から任意の多次元の度数表やクロス集計表を作り、変数値の分布なども表示できる。また、各種の検定や関連性の計算値を求めることも可能となっている。

MEANSプロシジャ概要

- MEANSプロシジャは、数値変数に対する単変量要約統計量を計算する。CLASSステートメントを指定しないと、入力データセットのすべてのオブザベーションを対象に統計量が計算される。またOUTPUTステートメントを使うと、統計量をSASデータセットに出力することができる。

The Power to Know.

# MEANSプロシジャの構文

```
PROC MEANS <options> <statistic-keyword-list>;
        VAR variable-list;
        CLASS variable-list;
        FREQ variable;
        WEIGHT variable;
        ID variable-list;
        BY variable-list;
        OUTPUT <OUT=SAS-data-set> <output-statistic-list>;
        <MAXID<(var-l<(id-list-l)><…var-n<id-list-n)>>)>=name-list>
        <MINID<(var-l<(id-list-l)><…var-n<id-list-n)>>)>=name-list>;
```

The Power to Know.

## サンプルプログラム1

```
proc means data=demo.sales;
     var amount;
run;
```

このサンプルプログラムでは、サンプルデータセットの変数名amount(売上高)の算出処理を行なっている。

The Power to Know.

## サンプルプログラム2

```
proc sort data=demo.sales;
     by product;
proc means data=demo.sales;
     by product;
     var amount;
run;
```

売上の集計を商品カテゴリごとに行なう場合には、グループ処理を行なう場合、BYステートメントを利用することができる。しかしながらBYステートメントを使ったグループ処理の場合は、BYステートメントで指定している変数でソートを行なっておく必要がある。

The Power to Know.

## サンプルプログラム３

```
proc means data=demo.sales;
     class product;
     var amount;
run;
```

CLASSステートメントに変数PRODCT(商品名)を指定してMEANSプロシジャを行なうと、元データをグループ処理したい変数で事前にソートをする必要もなく、またアウトプットの結果もCLASSステートメントで指定したカテゴリ(水準)がリスト形式で出てくるので見た目もわかりやすくなっている。

ただしCLASSステートメントを利用する場合は、すべてのグループをメモリに保持する必要があるので、大容量データセットを分析の場合はBY変数を利用したほうが良い場合もある

The Power to Know.

## サンプルプログラム４

```
proc means data=demo.sales;
     class dept product;
     var amount;
run;
```

Classステートメントに二つの変数を指定すると、指定した順番に集計が行われる。

The Power to Know.

## サンプルプログラム5

```
proc means data=demo.sales n sum range skewness
kurtosis;
      class product;
      var amount;
run;
```

MEANSプロシジャではPROC MEANSステートメントで特に指定をしなければ、N(件数)、MEAN(平均値)、STD(標準偏差)、MIN(最小値)、MAX(最大値)を算出するようになっているが、次のサンプルプログラムではその内容を変更してみる。

このサンプルプログラムではPROC MEANSステートメントでN(件数)、SUM(合計)、RANGE(範囲)、SKEWNESS(歪度)、KURTOSIS(尖度)を指定している。



## サンプルプログラム6

```
proc means data=demo.sales order=freq sum mean;
      class product;
      var amount;
run;
```

ORDERオプションにFREQを指定すると、度数が多い順番に出力される。

## サンプルプログラム7-1

```
proc means data=demo.sales;
      format date jdatewk8.;
      class date;
      var amount;
run;
```

FORMAT変換を行なうことにより、元のデータと違う切り口で集計することが可能となる。このデータでは日付は「YYMMDD10.」という年月日の日付形式になっているが、これを曜日ごとに変更して出力する場合は「jdatewk8.」などの曜日型に指定して出力する事が出来る。。



## サンプルプログラム7-2

```
proc format;
  value format a 0- 9999=' 1万円未満' 10000-49999=' 1万円～ 5万円未満'
        50000-99999=' 5万円～10万円未満' 100000-150000='10万円以上';
proc means data=demo.sales sum mean;
      format amount format_a.;
      class amount;
      var amount;
run;
```

また上記のように、「a」という名前のフォーマットを作成して、値段の区切りを設定して集計作業を行う事も可能となっている。

## サンプルプログラム８

```
proc means data=demo.sales;
      var amount;
      output out=out1;
run;
```

outputステートメントを指定すると、プロシジャの結果をデータセットとして出力が出来る。。
このデータセットは統計量を指定しないと、特殊変数_STAT_にN, MIN, MAX, MEAN, STDの統計量が出力される。また特殊変数_FREQ_は各分類レベルのオブザベーション数を表している。特殊変数_TYPE_はまたこの後に説明を行なう。

The Power to Know.

## サンプルプログラム９

```
proc means data=demo.sales;
      var amount;
      output out=out1 sum=gokei mean=heikin;
run;
```

サンプルプログラム８では、５つの統計量が出力されているが、上記のようにOUTPUTステートメントに指定することができる。 また複数の統計量を指定する事も出来る。上記の例では合計（SUM）を変数名GOKEIとして、平均（MEAN）を変数名HEIKINとして算出している。

The Power to Know.

# サンプルプログラム10

```
proc means data=demo.sales;
     class product;
     var amount;
     output out=out1 sum=gokei mean=heikin;
run;
```

出力データセットの１オブザベーション目には_TYPE_が「0」と「1」の2種類がある。この_TYPE_の値が「0」の場合はどのCLASS変数でも分類していないすべてのオブザベーションを利用している。この値が「1」の場合はCLASSステートメントで指定されたPRODUCT別に統計量を出力していることになる。CLASSステートメントで指定される変数が増えていけば、その分だけ_TYPT_の値も増えていき、各変数の統計量と各変数の水準を組み合わせた統計量が表示されるようになる。

The Power to Know.

# サンプルプログラム11

```
proc means data=demo.sales nway;
     class  shop dept product;
     var amount;
     output out=out1 sum=gokei mean=heikin;
run;
```

データウェアハウスなどから要約されたデータセットを利用する場合に、CLASSステートメントで指定された変数の組み合わせの統計量を出力させる場合にはNWAYオプションを指定して、_TYPE_の値が最大の値のオブザベーションのみを指定するようにする。

上記のように NWAYオプションを指定することにより、_TYPE_が「0」のものなどは出力されずに、CLASSステートメントで指定された、SHOP、DEPT、PRODUCTの各レベルの組み合わせられた結果が出力されている。

The Power to Know.


sas
e-Intelligence
The Power to Know.
The Power to Know.