本書では、IBM SPSS Modeler 18.2Jを使用しています。

IBM およびSPSSは、International Business Machines Corp.の登録商標です。

2 データのインポート

IBM SPSS Modelerでは、データベースやMicrosoft Excel形式、CSV形式などのさまざまなデータソースを読み込んで、必要なデータ加工や前処理を行い、分析やグラフ作成、モデリングを行うことができます。ここでは、Microsoft Excel形式とCSV形式のデータファイル、IBM SPSS Statistics形式とデータベースのテーブルを読み込む手順と注意点を確認します。

Contents

Keyword

Microsoft Excel / カンマ区切り / CSV / タブ区切り / データベースの読込み / ODBC /

§2.1.1　データのインポートの概要

IBM SPSS Modelerでは、**入力**ノードを使用することで、フラットファイル(カンマ区切りやタブ区切りなどのデータファイル)、IBM SPSS Statistics(.sav)、Microsoft Excel(.xls、.xlsx)、およびODBC 準拠のリレーショナルデータベースも含めたさまざまなフォーマットのデータソースを、インポートすることができます。また、ユーザー入力ノードを使用して、合成データを生成することもできます。

入力ノードは、**入力**パレットに含まれています。主な入力ノードの例は以下の通りです。

	**Excel**	Microsoft Excel（.xls、.xlsx)形式のデータをインポートします。ODBC データソースは不要です。
	**可変長ファイル**	フィールドがカンマやタブなどの区切り文字によって区切られているデータファイルをインポートします。
	**Statistics**	IBM SPSS Statistics(.sav)形式のデータファイルをインポートします。
	**データベース**	Microsoft SQL Server、DB2、Oracle などODBCを使用するデータベーステーブルをインポートします。
	**IBM Cognos Analytics**	IBM Cognos BI データベースからデータをインポートします。

Figure2.1.1　入力パレットに含まれる主な入力ノードの例

POINT

IBM SPSS Modelerに分析用のデータソースをインポートするためには、**入力**パレット内の**入力**ノードを使用します。

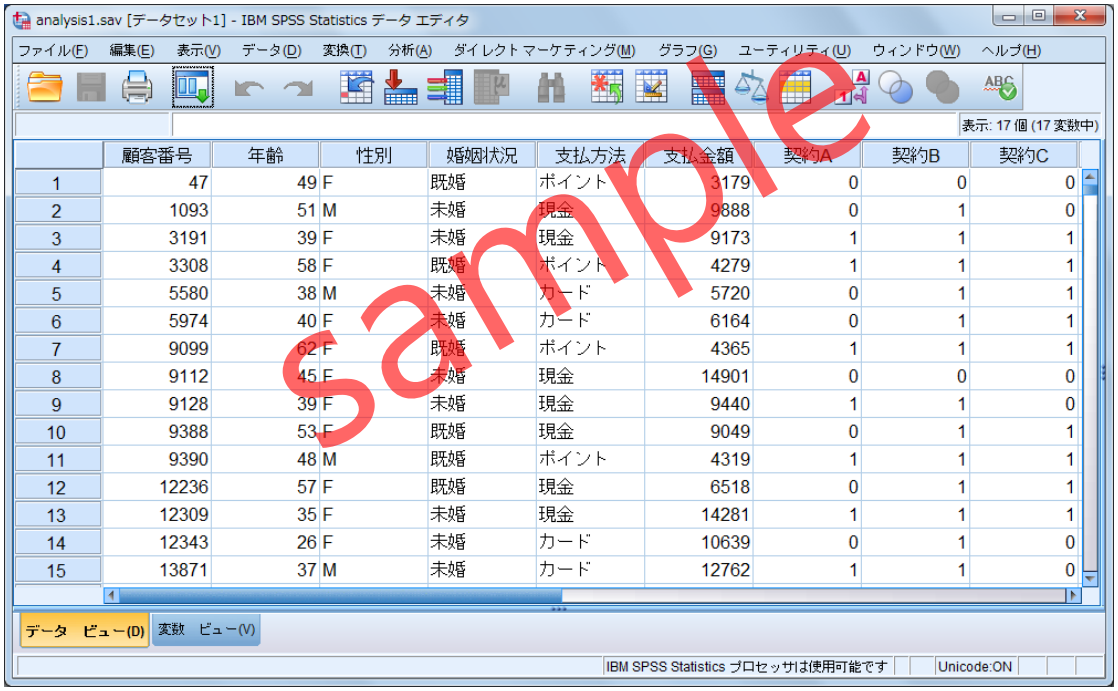
TIPS

レコード設定パレットの、**レコード結合**(フィールドの結合)ノードや**レコード追加**(レコードの結合)ノードを使用することで、異なるデータソースをIBM SPSS Modeler上で統合することができます。

§2.4.1　IBM SPSS Statistics形式のデータファイルの読込み

IBM SPSS Modelerでは、統計解析ソフトウェアの**IBM SPSS Statistics形式**のデータファイルを直接読み込むことができます。IBM SPSS Statisticsでは、変数の補足説明となる**変数ラベル**や、カテゴリ値の意味付けをする**値ラベル**の設定を行うことができ、IBM SPSS Modelerにも取り込むことができます。また、欠損値の設定などもIBM SPSS Statisticsで設定されている通りにIBM SPSS Modelerに読み込むことができます。

ここでは、以下のIBM SPSS Statistics形式のデータファイルをIBM SPSS Modelerに読み込みます。このデータファイルは、顧客の属性や利用サービスに関するデータが記録されています。フィールド(変数)は17個、8,632行のデータが含まれています。



	顧客番号	年齢	性別	婚姻状況	支払方法	支払金額	契約A	契約B	契約C
1	47	49	F	既婚	ポイント	3179	0	0	0
2	1093	51	M	未婚	現金	9888	0	1	0
3	3191	39	F	未婚	現金	9173	1	1	1
4	3308	58	F	既婚	ポイント	4279	1	1	1
5	5580	38	M	未婚	カード	5720	0	1	1
6	5974	40	F	未婚	カード	6164	0	1	1
7	9099	62	F	既婚	ポイント	4365	1	1	1
8	9112	45	F	未婚	現金	14901	0	0	0
9	9128	39	F	未婚	現金	9440	1	1	0
10	9388	53	F	既婚	現金	9049	0	1	1
11	9390	48	M	既婚	ポイント	4319	1	1	1
12	12236	57	F	既婚	現金	6518	0	1	1
13	12309	35	F	未婚	現金	14281	1	1	1
14	12343	26	F	未婚	カード	10639	0	1	0
15	13871	37	M	未婚	カード	12762	1	1	0

Figure2.4.1　analysis1.savファイル

列にフィールド(変数)、行にレコードが含まれています。また、Excelと同じように1行目には変数名を入力しておき、分析用のデータ値は2行目以降に入力しておきます。1列1変数、1行1ケースの形式になっていない場合、IBM SPSS Modelerでは正しくデータを読み込むことができません。

操作手順

1. **入力**パレットをクリックします。

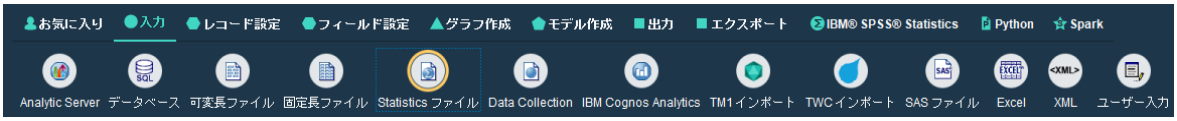


Figure2.4.2　入力パレットのStatisticsファイルノード

操作手順

2. **可変長ファイル**ノードを選択し、ストリームキャンバスに挿入します。

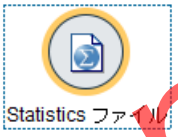


Figure2.4.3　ストリームキャンバスに配置したStatisticsファイルノード

ストリームキャンバスに挿入したノードは任意の場所に配置することができます。ノードの編集画面を利用することで、読み込むファイルの選択や詳細の設定を行います。

POINT

Statisticsファイルノードは、IBM SPSS Statistics形式のデータファイルのインポートに使用します。

操作手順

3. **Statistics**ノードをダブルクリックして編集画面を開きます。

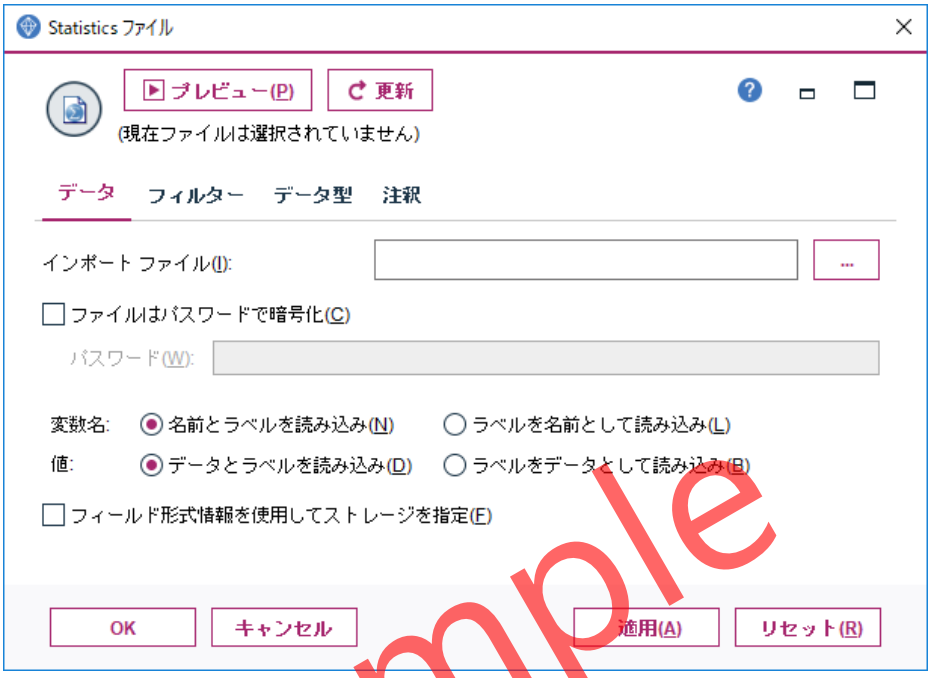


Figure2.4.4　Statisticsファイルノードの編集画面

データタブを利用して、読込むファイルの指定を行います。

インポートファイルボックスに、読込みが必要なデータファイル名を指定します。

変数名では、フィールド(変数)の名前とラベルの両方を読み込むか、ラベルのみを読み込むかを選択します。また、値では、カテゴリ値と値ラベルの両方を読み込むか、値ラベルのみを読み込むかを選択します。この例では、デフォルトのまま進めます。

操作手順

4. **インポートファイル**の**ファイルを探す**ボタンをクリックします。
5. ファイルの一覧から**analysis1.sav**を選択します。

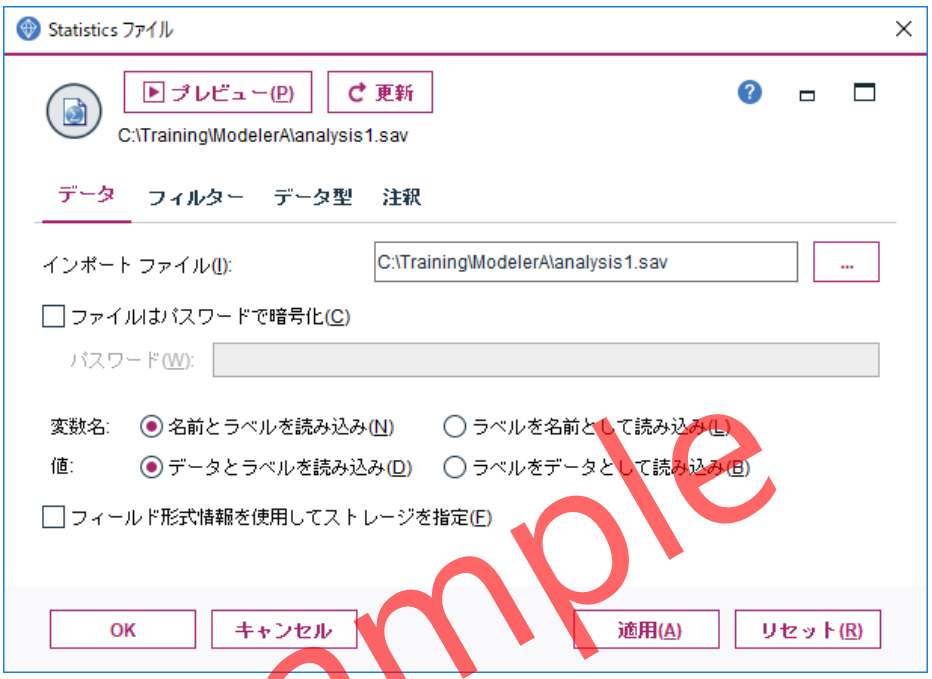


Figure2.4.5　データファイルの選択が完了したStatisticsファイルノードのデータタブ

操作手順

6. **変数名**の**名前とラベルを読み込み**を選択します。
7. **値**の**データとラベルを読み込み**を選択します。

TIPS

IBM SPSS Statistics形式のデータファイルでは、変数名を補足するための**変数ラベル**と、カテゴリ値の意味付けをするための**値ラベル**を設定することができます。

操作手順

8. **データ型**タブをクリックします。

Figure2.4.6　Statisticsノードのデータ型タブ

データ型タブでは、各フィールドの**尺度**や**値**、**欠損値**や**ロール**の設定を行うことができます。IBM SPSS Statisticsで設定済みの定義は、自動的にセットされます。これらの設定は、集計やグラフ作成、モデリングに影響するため非常に重要な設定ですが、尺度の設定等は次の章で確認しますので、ここでは設定しません。

TIPS

IBM SPSS Statisticsでは、変数ビューを利用して、変数の測定の尺度や値ラベル、ユーザー欠損値の指定などを行っておくことができ、これらの情報をIBM SPSS Modelerに読み込むことができます。

操作手順

9. **OK**ボタンをクリックします。

Figure2.4.7　ファイルを設定したStatisticsファイルノード

ストリームキャンバスに表示されるノード名が、設定したファイル名に変わります。次に、**テーブル**ノードを利用して、データをIBM SPSS Modelerに読み込みます。

操作手順

10. ストリームキャンバスの**analysis1.sav**を選択しておきます。
11. **出力**パレットをクリックします。

Figure2.4.8　出力パレットのテーブルノード

操作手順

12. **テーブル**ノードをダブルクリックします。
13. ツールバーの**選択内容を実行**ボタンをクリックします。

	顧客番号	年齢	性別	婚姻状況	支払方法	支払金額	契約A	契約B
1	47.000	49.000	F	既婚	ポイント	3179.000	0.000	0.000
2	1093.000	51.000	M	未婚	現金	9888.000	0.000	1.000
3	3191.000	39.000	F	未婚	現金	9173.000	1.000	1.000
4	3308.000	58.000	F	既婚	ポイント	4279.000	1.000	1.000
5	5580.000	38.000	M	未婚	カード	5720.000	0.000	1.000
6	5974.000	40.000	F	未婚	カード	6164.000	0.000	1.000
7	9099.000	62.000	F	既婚	ポイント	4365.000	1.000	1.000
8	9112.000	45.000	F	未婚	現金	14901.000	0.000	0.000
9	9128.000	39.000	F	未婚	現金	9440.000	1.000	1.000
10	9388.000	53.000	F	既婚	現金	9049.000	0.000	1.000
11	9390.000	48.000	M	既婚	ポイント	4319.000	1.000	1.000
12	12236.000	57.000	F	既婚	現金	6518.000	0.000	1.000
13	12309.000	35.000	F	未婚	現金	14281.000	1.000	1.000
14	12343.000	26.000	F	未婚	カード	10639.000	0.000	1.000
15	13871.000	37.000	M	未婚	カード	12762.000	1.000	1.000
16	14211.000	35.000	M	未婚	現金	4644.000	0.000	0.000
17	14921.000	44.000	M	未婚	現金	5427.000	0.000	1.000
18	15068.000	40.000	F	既婚	カード	20065.000	0.000	1.000
19	15167.000	63.000	F	既婚	ポイント	6470.000	1.000	1.000
20	15252.000	50.000	F	既婚	現金	11751.000	1.000	1.000

Figure2.4.9　テーブルに出力されたanalysis1.csvのデータ

データが読み込まれ、テーブルに出力されます。タイトルバーの表示から、17フィールド、8,632レコードが含まれていることが分かります。

操作手順

14. テーブルを閉じます。

4 フィールドの要約

フィールドは、**変数** variables とも呼ばれ、年齢や性別、購入商品やアンケートの設問など、個々のデータを記録する項目です。個々のフィールドに注目してフィールドの持つ情報を要約することを、**1フィールドの要約**と呼びます。

フィールドの要約を行う場合、各フィールドの**尺度**の設定が重要です。1フィールドの要約は、クロス集計表や相関係数を計算する2フィールドの関係性の分析、ディシジョンツリーやロジスティック回帰分析などのようなモデリングの基礎にもなる重要な処理です。

Contents

Keyword

フィールドの要約 / 要約統計量 / 平均値 / 中央値 / 最頻値 / 最小値 / 最大値 / 範囲 / 標準偏差 / 分散 / 四分位範囲 / パーセンテージ / 記述統計ノード / データ検査ノード / ヒストグラム / 棒グラフ / 欠損値 / 欠損値選択ノード /

§4.1.1　フィールドの要約の概要

フィールドが持つデータは、さまざまな方法で要約することができます。最も基本となるのは**1フィールドの要約**です。基本的に、連続型フィールドは平均値をはじめとする要約統計量によって要約し、カテゴリ型フィールドはパーセンテージなどによって要約します。

	顧客番号	年齢	性別	婚姻状況	支払方法	支払金額	契約A	契約B
1	47	49	F	既婚	ポイント	3179	0	0
2	1093	51	M	未婚	現金	9888	0	1
3	3191	39	F	未婚	現金	9173	1	1
4	3308	58	F	既婚	ポイント	4279	1	1
5	5580	38	M	未婚	カード	5720	0	1
6	5974	40	F	未婚	カード	6164	0	1
7	9099	62	F	既婚	ポイント	4365	1	1
8	9112	45	F	未婚	現金	14901	0	0
9	9128	39	F	未婚	現金	9440	1	1
10	9388	53	F	既婚	現金	9049	0	1
11	9390	48	M	既婚	ポイント	4319	1	1
12	12236	57	F	既婚	現金	6518	0	1
13	12309	35	F	未婚	現金	14281	1	1
14	12343	26	F	未婚	カード	10639	0	1
15	13871	37	M	未婚	カード	12762	1	1
16	14211	35	M	未婚	現金	4644	0	0
17	14921	44	M	未婚	現金	5427	0	1
18	15068	40	F	既婚	カード	20065	0	1
19	15167	53	F	既婚	ポイント	6470	1	1
20	15252	50	F	既婚	現金	11751	1	1

Figure4.1.1　テーブルノードでの生データの表示

データ分析においては、できるだけ生データを確認することが重要です。しかし、フィールドやレコードの数が大きくなるほど、生データから何らかの傾向や特徴を見出すことは困難になります。そこで、さまざまな統計量を用いてデータの要約を図ります。

フィールドの要約の最も基本となるのが、1フィールドの要約です。さらに、2フィールドの関係性、多数のフィールドに基づくモデリングと拡張されていきます。

§4.1.2　要約統計量

要約統計量 summary statistics は、フィールドの持つ情報を要約するための指標です。記述統計量とも呼ばれます。要約統計量は、使用するフィールドの尺度によって、適切に使い分ける必要があります。フィールドの要約では、**中心傾向**と**散らばり**の2つの視点が必要です。

POINT

中心傾向を示す要約統計量には、**平均値**、**中央値**、**最頻値**があります。

平均値 mean は、フィールドのすべてのデータ値を足して、データ数で割ることによって得られる値です。もっともよく利用される中心傾向の要約統計量ですが、外れ値や分布の歪みの影響を受けやすい特性を持ちますので、使用する際はデータの分布を確認することが重要です。

中央値 median は、データ値を大きさの順番に並び替えたときに、ちょうど真ん中に位置する値です。中央値より大きい値と小さい値がちょうど同数になります。データ数が偶数の場合は、真ん中の2つの値の平均値が中央値になります。中央値は外れ値の影響を受けません。

最頻値 mode は、もっとも頻度の高い値です。多数決的な要約統計量です。

それぞれの測定の尺度で使用することができる、要約統計量は以下のとおりです。

		名義型	順序型	連続型
平均値	Mean	×	△	○
中央値	Median	×	○	○
最頻値	Mode	○	○	○

Table4.1.1　中心傾向の要約統計量と測定の尺度

TIPS

順序型は、原則として平均値を使用することは適切ではありません。ただし、アンケート評価のようなフィールドで、**5段階**以上の尺度がある場合は、便宜上等間隔とみなして統計処理を行うケースが多いです。この場合は、データを数値として入力しておき、尺度を順序型ではなく連続型として設定しておく必要があります。

POINT

変数の**散らばり**を示す要約統計量には、**最小値**、**最大値**、**範囲**、**四分位範囲**、**分散**、**標準偏差**があります。

最小値 minimum と**最大値** maximum は、それぞれのデータ値のうち最も大きな値と最も小さな値を意味します。最大値と最小値の差が**範囲** range です。

四分位範囲 inter quartile range, **IQR** は、データを並び替えてサンプルの25%ずつが含まれるように4等分したときの真ん中の50%の範囲(75パーセンタイル値と25パーセンタイル値の差分)です。

分散 variance は、平均値に対するデータのばらつきを意味し、分散の平方根が**標準偏差** Standard Deviation, **SD**です。分散はデータ値と平均値の差(偏差)を2乗して計算されるため、元のデータと単位が異なりますが、標準偏差は分散の平方根となるため、元のデータと同じ単位で解釈できます。

それぞれの測定の尺度で使用することができる、要約統計量は以下のとおりです。

		名義型	順序型	連続型
最小値	Minimum	×	○	○
最大値	Maximum	×	○	○
範囲	Range	×	○	○
四分位範囲	Inter Quartile Range	×	○	○
分散	Variance	×	△	○
標準偏差	Standard Deviation	×	△	○

Table4.1.2　散らばりの要約統計量と測定の尺度

POINT

分散はデータ値と平均値の差(偏差)を2乗して計算されるため、元のデータと単位が異なりますが、**標準偏差**は分散の平方根となるため、元のデータと同じ単位で解釈できます。

§4.1.3　ストリームの確認

1フィールドの要約の手順を確認するために、サンプルのストリームファイル**Chapter4.str**を読み込みます。この章では、**記述統計**ノードや**データ検査**ノードによる分析を行います。

操作手順

1. **ファイル**メニュー ＞ **ストリームを開く**を選択します。
2. **/ ModelerA / Chapter4.str**を選択します。

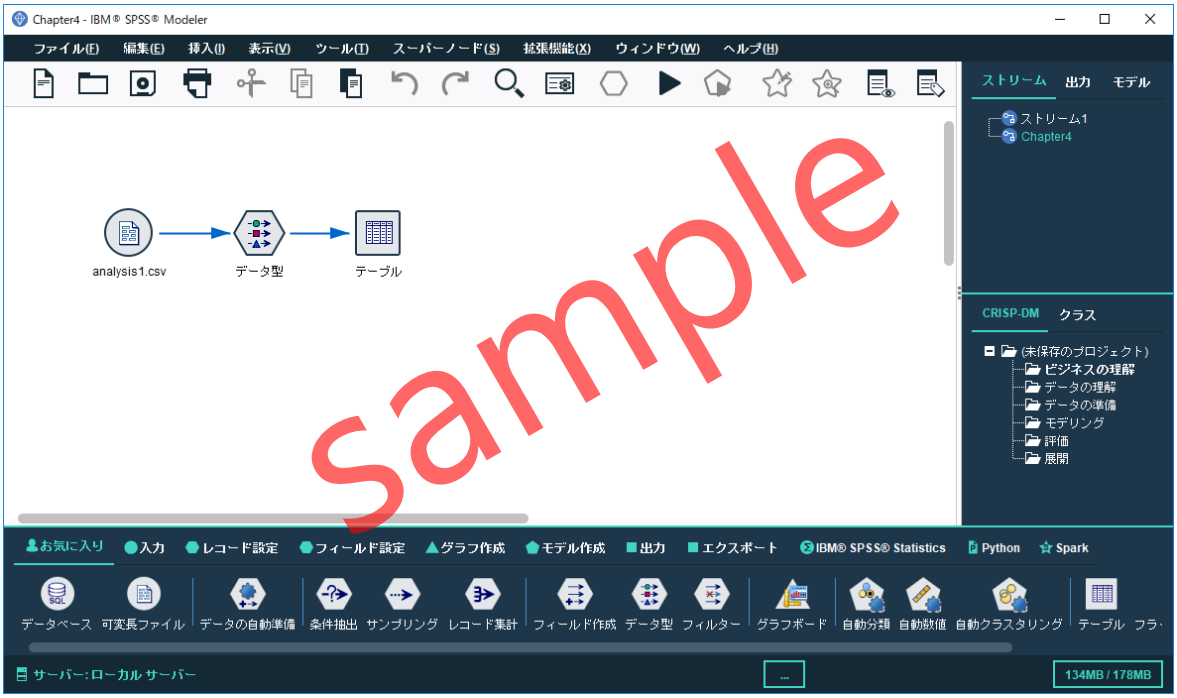


Figure4.1.2　Chapter4.strファイルのストリーム

このファイルには、カンマ区切りのデータソース**analysis1.csv**を読み取り、テーブルに出力する簡単なストリームが含まれています。

入力ノードと**テーブル**ノードの間に、**データ型**ノードが挿入されており、尺度やロールの設定が行われています。

§4.3.1　データ検査の実行

データ検査ノードを用いると、各フィールドの要約情報を出力することができます。フィールドの尺度に応じたグラフとして、カテゴリ型フィールドでは棒グラフ、連続型フィールドではヒストグラムが自動的に作成されます。

データ検査ノードは、**出力**パレットに含まれています。

操作手順

1. **出力パレット**を開きます。

Figure4.3.1　出力パレットのデータ検査ノード

操作手順

2. **データ検査**ノードを選択ストリームキャンバスに挿入し、データ型ノードからリンクします。

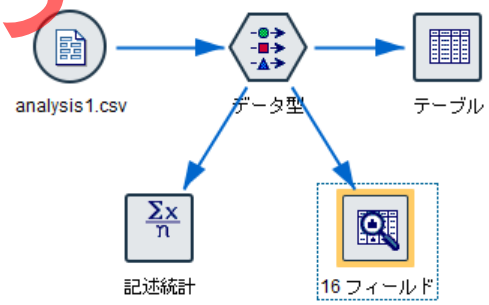


Figure4.3.2　データ型ノードからリンクしたデータ検査ノード

データ型ノードから16個のフィールを受け取っていることが分かります。

操作手順

3. **データ検査**ノードをダブルクリックして編集画面を開きます。

Figure4.3.3　データ検査ノードの編集画面

設定タブを利用して、データ検査を行うフィールドを指定します。**デフォルト**では、上流から受け取ったフィールドが使用されます。**ユーザー設定フィールドを使用**に切り替えると、個別にフィールドの指定を行うことができます。

表示では、**グラフ作成**と**基本統計量**が指定されています。基本統計量では、連続型フィールドについて平均値、標準偏差、最小値、最大値、歪度を計算して出力します。**高度な統計**を選択すると、合計、範囲、平均値の標準誤差、尖度などが追加で出力されます。中央値と最頻値はデフォルトでは出力されませんので、使用する場合は選択する必要がありますが、データの並べ替えが行われるため、パフォーマンスが低速になる可能性があります。

操作手順

4. **実行**ボタンをクリックします。

§4.3.2　データ検査の結果の解釈　ー連続型フィールドー

出力されるデータ検査の結果を確認します。この例では、16個のフィールドの情報が要約されており、連続型フィールドは最小値、最大値、平均値、標準偏差などによってまとめられ、フラグ型や名義型などのカテゴリフィールドはカテゴリ数がまとめられています。

フィールド	サンプル グラフ	尺度	最小値	最大値	平均値	標準 偏差	歪度	カテゴリ数	有効
年齢		連続型	18	63	43.496	10.858	-0.071	--	8632
性別		フラグ型	--	--	--	--	--	2	8632
婚姻状況		フラグ型	--	--	--	--	--	2	8632
支払方法		名義型	--	--	--	--	--	3	8632
支払金額		連続型	214	30363	9421.189	4666.701	0.718	--	8632
契約A		フラグ型	0	1	--	--	--	2	8632
契約B		フラグ型	0	1	--	--	--	2	8632
契約C		フラグ型	0	1	--	--	--	2	8632

Figure4.3.4　データ検査ノードの実行結果

TIPS

フィールドのロールに**対象**がセットされていると、対象フィールドの値ごとにオーバーレイされたグラフが出力されます。

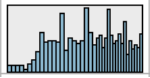
フィールド	サンプル グラフ	尺度	最小値	最大値	平均値	標準 偏差	歪度	カテゴリ数	有効
年齢		連続型	18	63	43.496	10.858	-0.071	--	8632

Figure4.3.5　年齢フィールドのデータ検査結果

年齢は連続型のフィールドです。最小値=**18**、最大値=**63**であり、顧客の年齢は18歳から63歳の間です。平均値=**43.496**であり、顧客の平均年齢は約43歳であることが分かります。また、標準偏差=**10.858**であり、平均年齢に対して約11歳のばらつきを持つと解釈することができます。これらは、記述統計ノードで確認した結果と同じです。

フィールドが正規分布に近い場合、平均値±1標準偏差内におよそ60～70%のデータが含まれることが示唆されます。つまり、43±11の計算から、顧客の60～70%は、32歳～54歳の範囲に含まれそうです。

POINT

フィールドが正規分布に近い場合、平均値±**1**標準偏差内におよそ60～70%のデータが含まれます。

TIPS

フィールドが正規分布に近い場合、平均値±**2**標準偏差内におよそ95%のデータが含まれます。平均値±**3**標準偏差内におよそ99%のデータが含まれます。

サンプルグラフをダブルクリックすると、拡大して詳細を確認することができます。

操作手順

1. **年齢**のサンプルグラフをダブルクリックします。

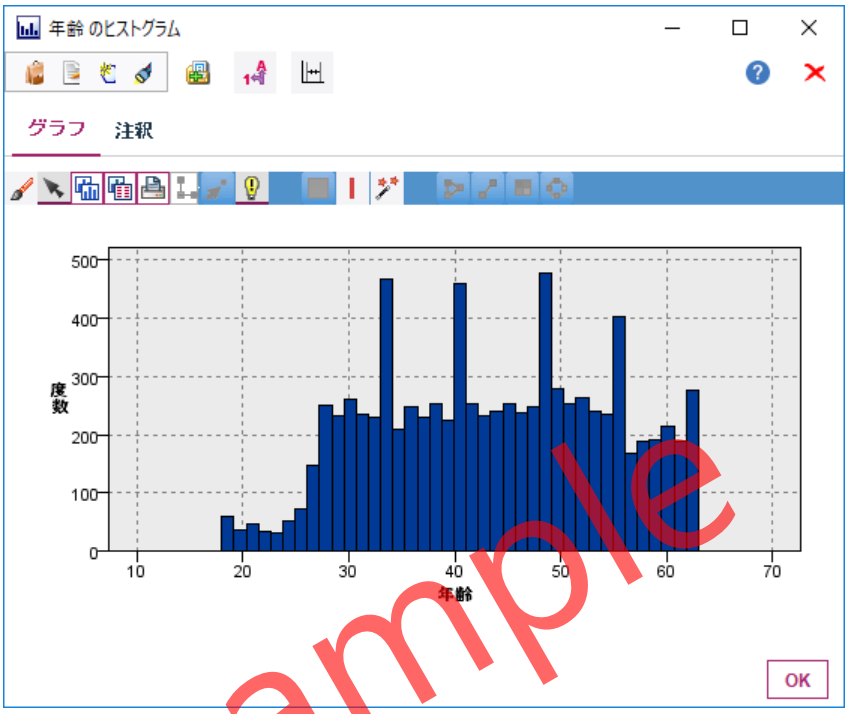


Figure4.3.6　年齢のヒストグラム

連続型フィールドを要約するグラフとして**ヒストグラム**が表示されます。ヒストグラムの横軸には連続型フィールド、縦軸に度数が表示されます。

POINT

連続型フィールドを要約するグラフは**ヒストグラム**です。

操作手順

2. **OK**ボタンをクリックして、ヒストグラムを閉じます。