§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。

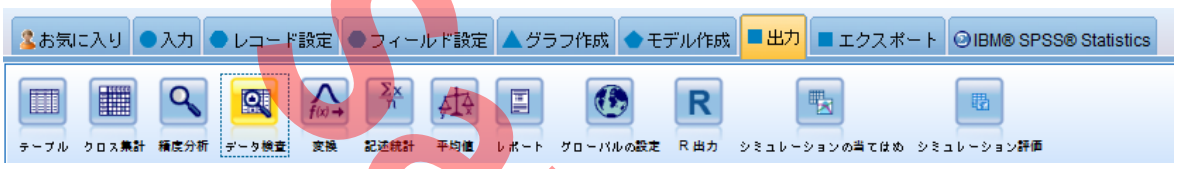


Figure4.3.1　出力パレットのデータ検査ノード

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。

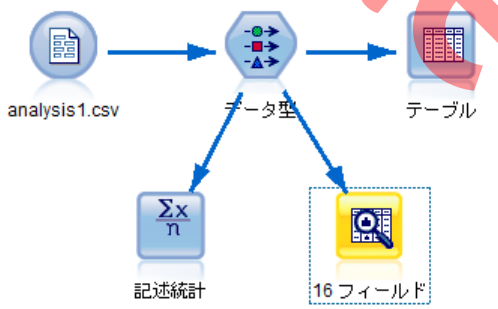


Figure4.3.2　データ型ノードからリンクしたデータ検査ノード

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。

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。

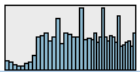
フィールド	サンプル グラフ	尺度	最小値	最大値	平均値	標準偏差	歪度	カテゴリー数	有効
年齢		連続型	18	63	43.496	10.858	-0.071	--	8632

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。

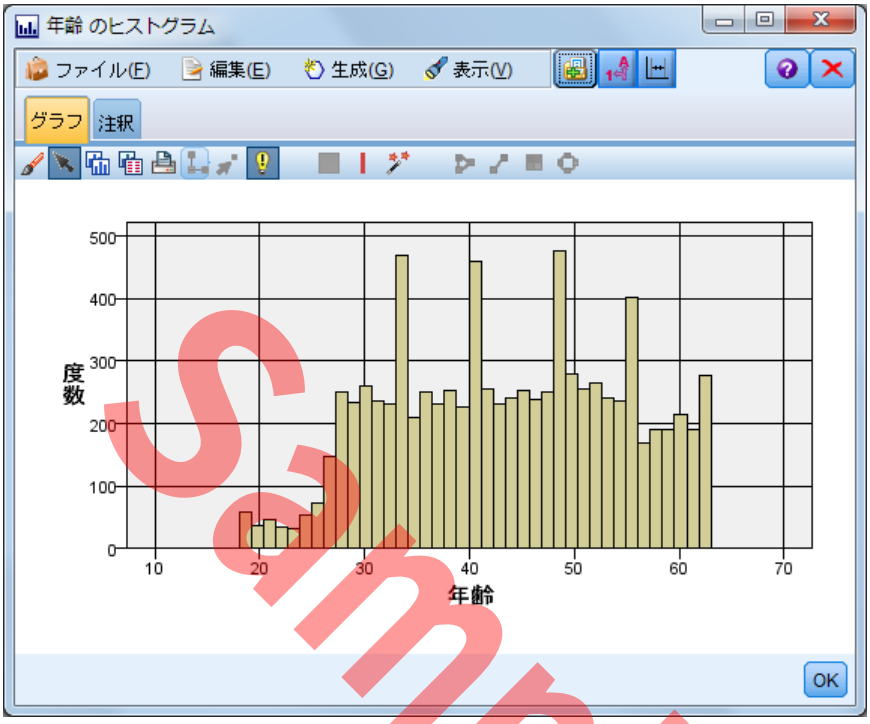


Figure4.3.6　年齢のヒストグラム

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。

§4.3.3　データ検査の結果の解釈　－カテゴリ型フィールド－

次に、カテゴリ型フィールドのデータ検査結果を確認します。カテゴリ型フィールドでは、平均値や標準偏差などの要約統計量の使用は適切ではないため、カテゴリ数による要約が行われます。

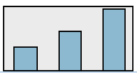
フィールド	サンプル グラフ	尺度	最小値	最大値	平均値	標準偏差	歪度	カテゴリー数	有効
A 支払方法		名義型	--	--	--	--	--	3	8632

Figure4.3.7　データ検査ノードによるカテゴリ型フィールドの要約結果の例

支払方法は、**名義型**のフィールドです。このフィールドに含まれるカテゴリー数は3個であることが分かります。カテゴリ型フィールドの場合、連続型フィールドのような最小値、最大値、平均値、標準偏差は計算されません。カテゴリ型フィールドの場合は、各カテゴリの度数やパーセンテージを確認します。パーセンテージはサンプルグラフで確認することができます。

POINT

カテゴリ型フィールドには、性別や地域、支払方法などカテゴリのデータ値が含まれます。したがって、このタイプのフィールドを使用する場合、最小値、最大値、平均値、標準偏差は計算されません。

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. **支払方法**のサンプルグラフをダブルクリックします。

Figure4.3.8　支払方法の棒グラフのテーブル表示

棒グラフが表示され、**テーブル**タブでは各カテゴリの%と度数を確認することができます。この例では、**カード**が1625レコードで18.83%、**ポイント**が2662レコードで30.84%、**現金**が4325レコードで50.34%です。支払方法として最も多いのは現金であることが分かります。

TIPS

割合や**%**、**度数**の列名をクリックするごとに、値に基づいて表示の**昇順**と**降順**を切り替えることができます。

操作手順

2. **グラフ**タブをクリックします。

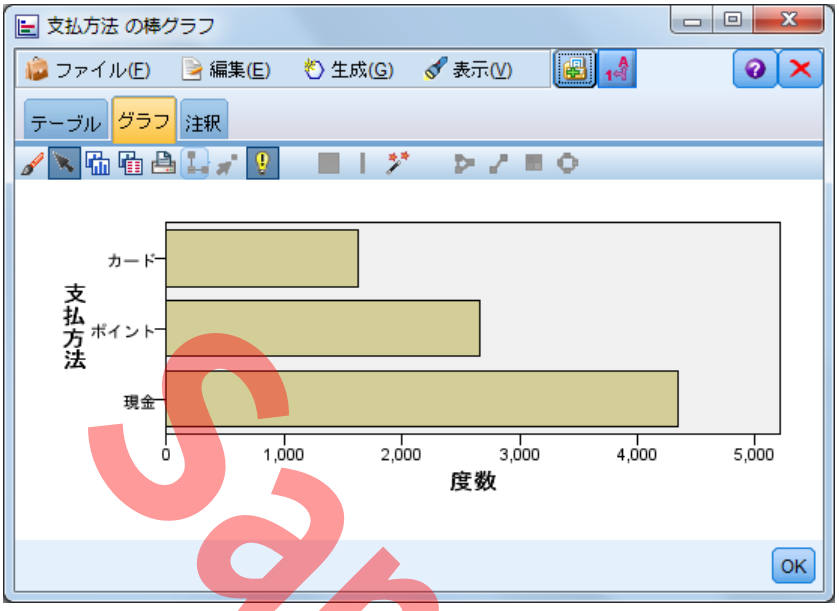


Figure4.3.9　棒グラフのグラフタブ表示

グラフタブの表示はグラフメインとなります。この表示は、編集機能を用いて、グラフ内に含めるカテゴリの切替えや並び替えなどを行うことができます。

POINT

カテゴリ型フィールドを要約するグラフは、棒グラフです。

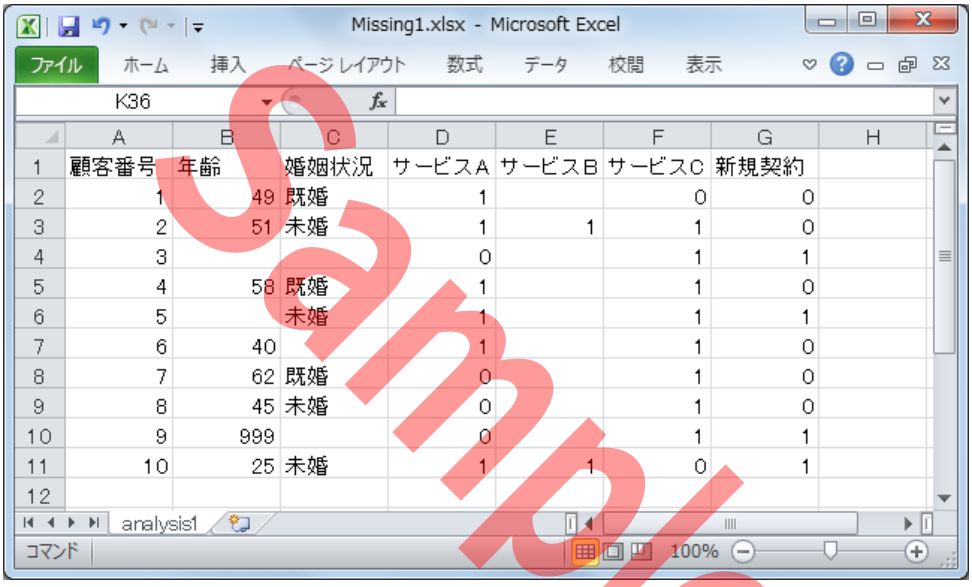
TIPS

グラフタブでは、**編集モード**に切り替えることでグラフの編集を行うことができます。グラフ編集の具体的な手順は、Chapter5で確認します。

§4.4.1　欠損値の概要

データソースに含まれるフィールドには、さまざまな理由で欠損値が発生することがあります。欠損値は、アンケートの無回答やデータ値の不明などによる未入力をあらわし、原則として分析から除外されます。

ここでは、以下のExcel形式のデータファイル**Missing1.xlsx**をIBM SPSS Modelerに読み込みます。このデータファイルには、7個のフィールド(変数)、10行のデータが含まれています。



	A	B	C	D	E	F	G	H
1	顧客番号	年齢	婚姻状況	サービスA	サービスB	サービスC	新規契約	
2	1	49	既婚	1		0	0	
3	2	51	未婚	1	1	1	0	
4	3			0		1	1	
5	4	58	既婚	1		1	0	
6	5		未婚	1		1	1	
7	6	40		1		1	0	
8	7	62	既婚	0		1	0	
9	8	45	未婚	0		1	0	
10	9	999		0		1	1	
11	10	25	未婚	1	1	0	1	
12								

Figure4.4.1　欠損値を含むデータソースの例

IBM SPSS Modelerで認識される欠損値には以下の種類があります。

ヌル値	数値型フィールドに含まれる欠損値(未入力)データです。
空文字列	文字型フィールドに含まれる欠損値(未入力)データです。
空白文字	文字型フィールドに含まれるタブやスペースです。
空白値	ユーザー指定の欠損値です。

Table4.4.1　IBM SPSS Modelerで認識される欠損値