

データ分析の基礎-1

2022年10月6日



本日の内容

- 1 データの代表値
平均・バラツキ、中央値
- 2 データの視覚化
箱ひげ図、ヒストグラム
- 3 正規分布、t 分布
- 4 t 検定
1標本、2標本

データの代表値

平均（算術平均）

（例）英語の研修前後の成績（20人）

	A君の得点	全体の平均点	得点-平均点
研修前	70	58.3	+11.7
研修後	75	58.3	+13.7

A君の成績の評価は？

平均の信頼性？

平均の信頼性

(例) 暑い日に暑い場所で待ち合わせをした。
いつも遅れてくる人が何分後に来るのか予測。

<過去10回の遅刻データ (分)>

平均

①	21	46	8	28	19	34	13	33	19	31	25.2分
②	25	26	23	24	26	27	26	25	24	26	25.2分
③	21	22	21	23	22	23	22	21	23	54	25.2分

①のデータは「バラツキ」が大きい。

「バラツキ」の大きいデータの平均は信頼できない。

③のデータは「外れ値」を含む。

「外れ値」を含むデータの平均は信頼できない。

中央値 (MEDIAN)

230	287	252	313	825	331	292	851	229
-----	-----	-----	-----	-----	-----	-----	-----	-----

平均値 = 401.1

平均値は極端な値の影響を受け易い



1	2	3	4	5	6	7	8	9
---	---	---	---	---	---	---	---	---

229	230	252	287	292	313	331	825	851
-----	-----	-----	-----	-----	-----	-----	-----	-----

大きさの順番に並び替えて、中央に位置する値を代表値

データの視覚化

データを分析するには全体を把握することが重要。

(1) 各データ群の把握

①外れ値の有無の確認

②分布の偏り

箱ひげ図、ヒストグラム

(2) 2群以上のデータ群の関係の把握

散布図

外れ値

1	2	3	4	5	6	7	8	9
229	230	252	287	292	313	331	825	851

Q1 第1四分位数 (25%) : 252

Q2 第2四分位数 (50%) : 292

Q3 第3四分位数 (75%) : 331

IQR 四分位範囲 $Q3 - Q1 = 331 - 252 = 79$

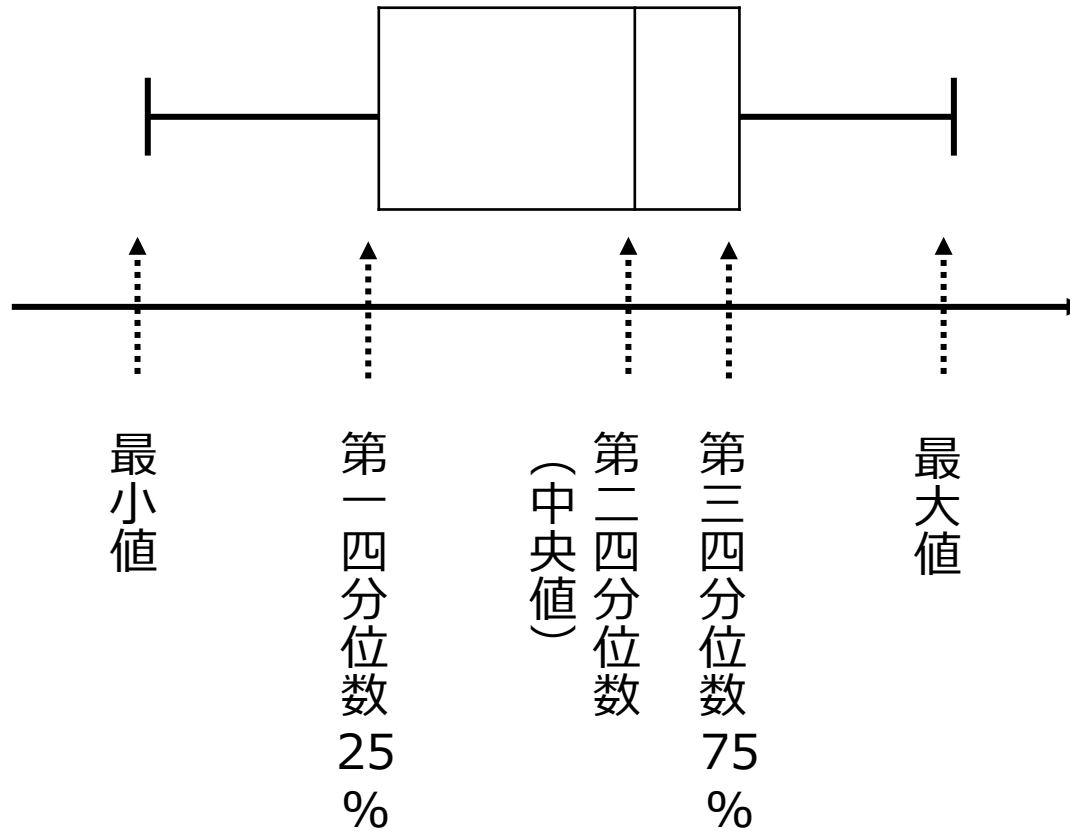
$Q1 - 1.5 \times IQR$ 以下 $Q3 + 1.5 \times IQR$ 以上

$252 - 1.5 \times 79 = 133.5$ $331 + 1.5 \times 79 = 449.5$

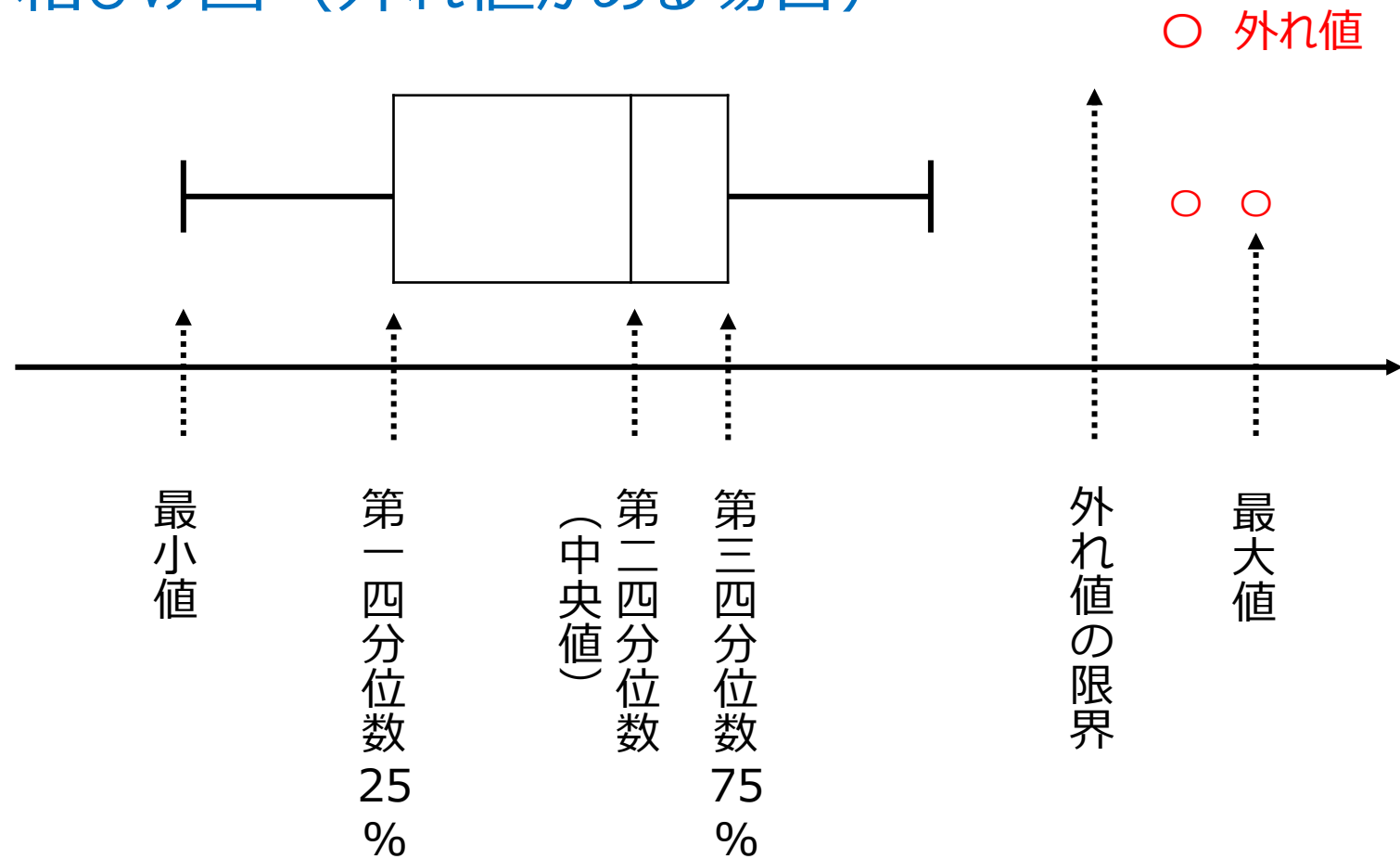
外れ値 : 133.5以下 449.5以上

箱ひげ図

(外れ値がない場合)



箱ひげ図（外れ値がある場合）



SAS Studio

- 1.SAS Studio にログインする。
- 2.SAS®Studio をクリックする。

The screenshot shows the SAS OnDemand for Academics Sign In page. At the top, there is a 'Notices' button. Below it, the text 'SAS® OnDemand for Academics Sign In' is displayed. There are two input fields: 'SAS Profile email address or user ID' and 'Password'. Below the password field, there is a checkbox for 'Accept the terms of service' and a link to 'use and conditions'. A tooltip for 'SAS® Studio' is overlaid on the bottom right, showing the SAS Studio icon, the text 'SAS® Studio', a description 'Write and run SAS code with a Web-based SAS development environment.', and an action 'Clear my saved tabs.'.

Notices

SAS® OnDemand for Academics
Sign In

SAS Profile email address or user ID

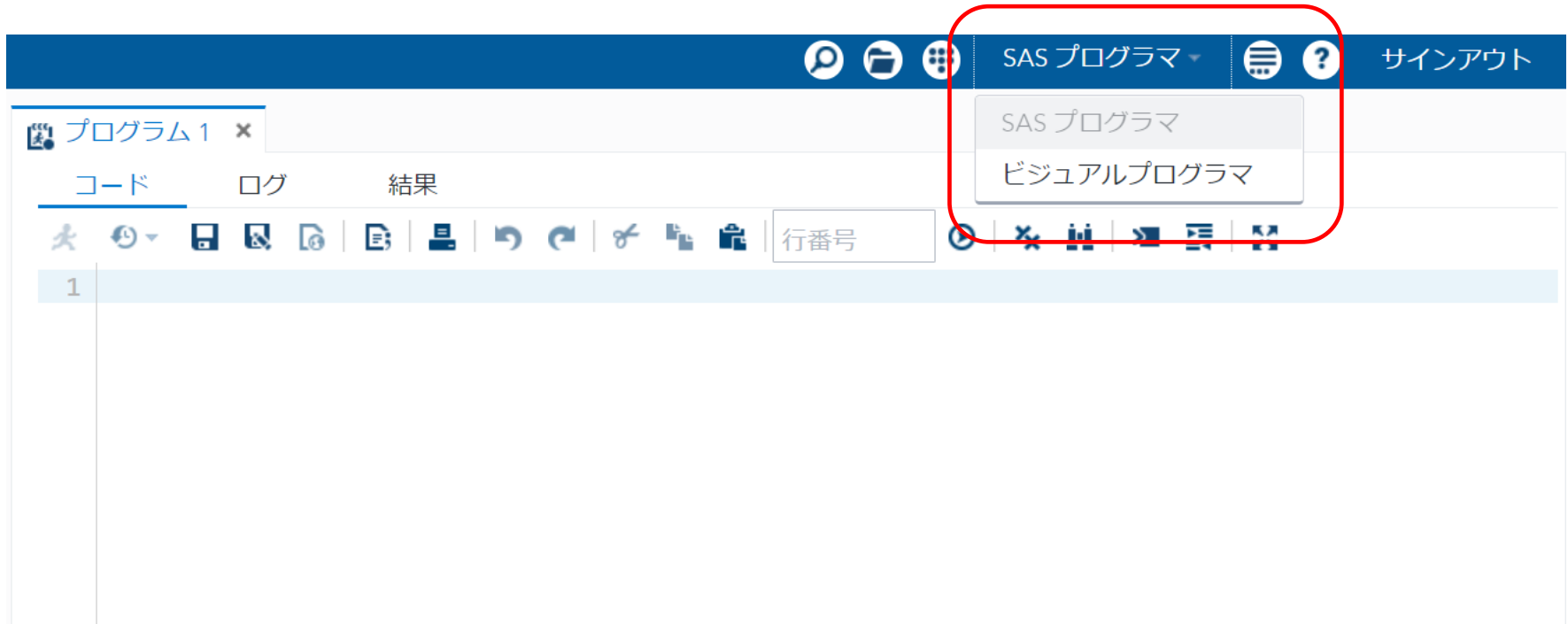
Password

☐ Accept the terms of service [use and conditions](#)

Applications Enrollments Courses

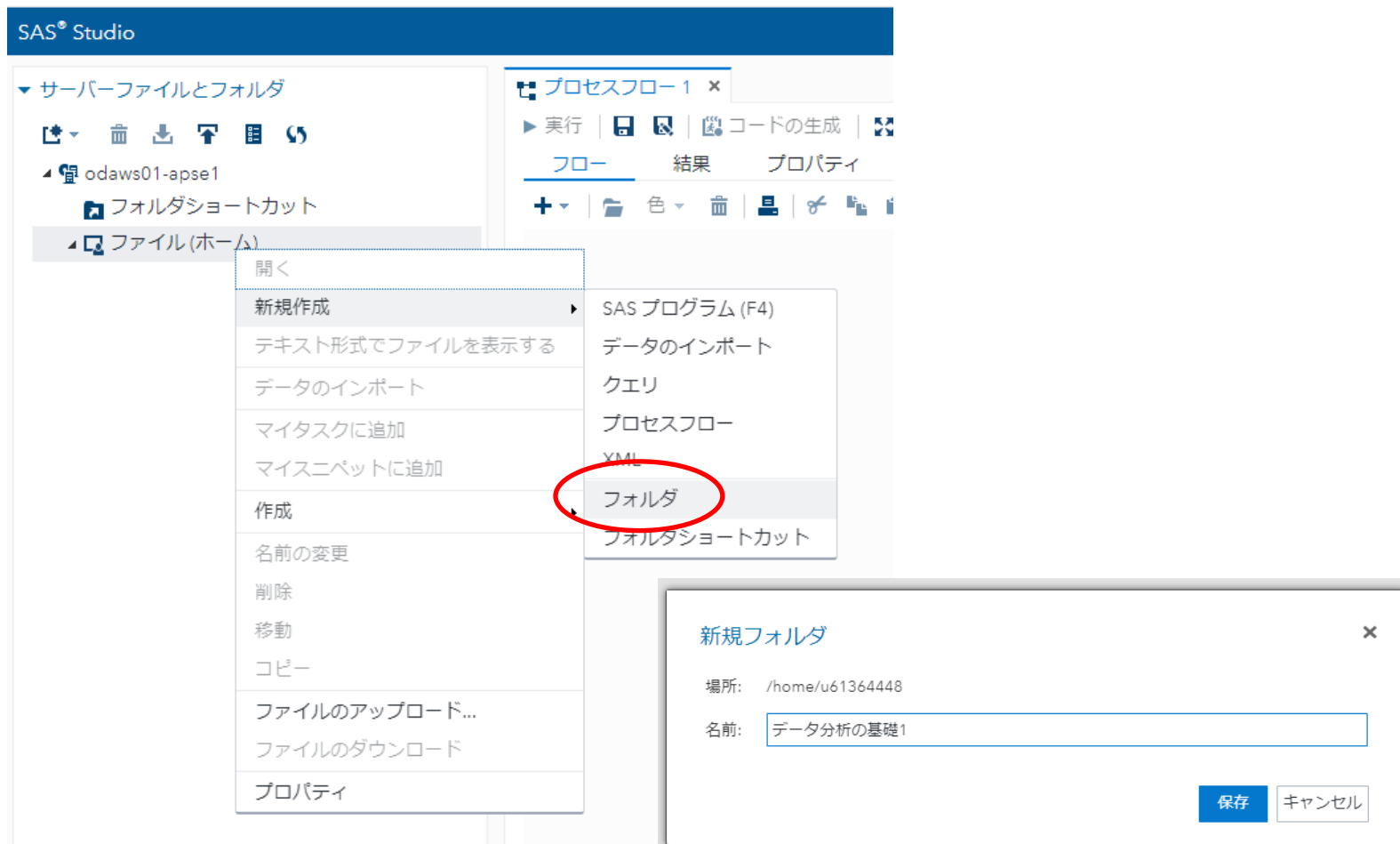
 **SAS® Studio**
Write and run SAS code with a Web-based SAS development environment.
Actions: [Clear my saved tabs.](#)

3. 「SASプログラマ」をクリックし、「ビジュアルプログラマ」を選択する。



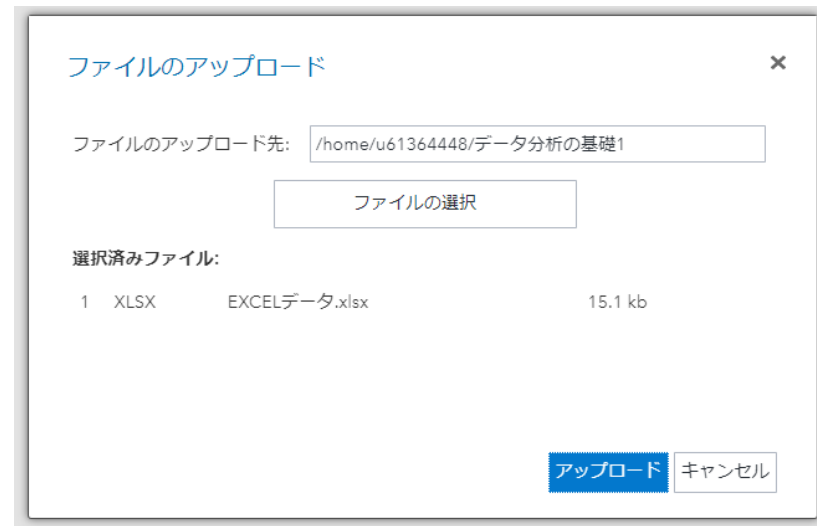
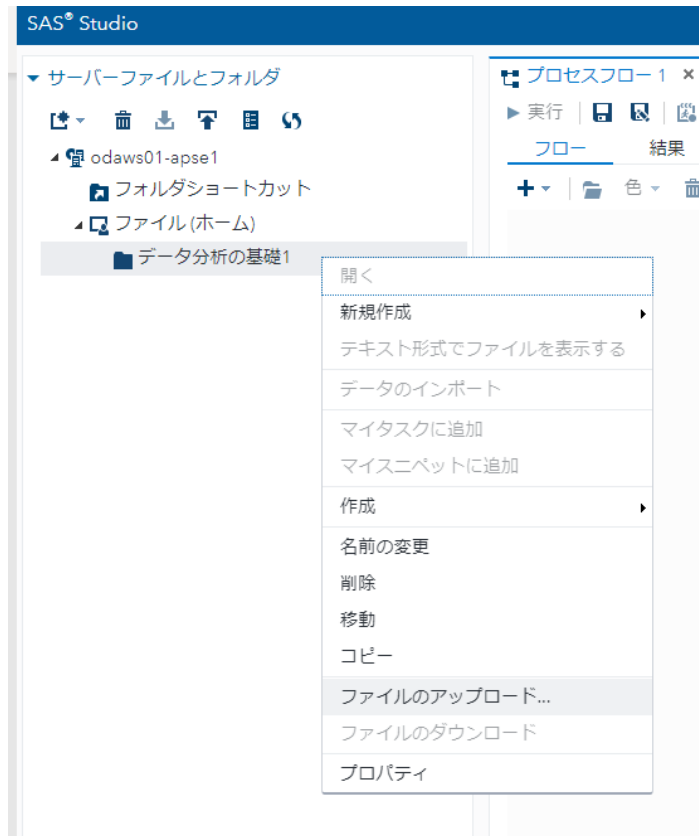
4. 新規フォルダを作成する。

「ファイル（ホーム）」を右クリック→「新規作成」→「フォルダ」をクリック、
「新規フォルダ名（データ分析の基礎1）」を入力し、「保存」をクリックする。

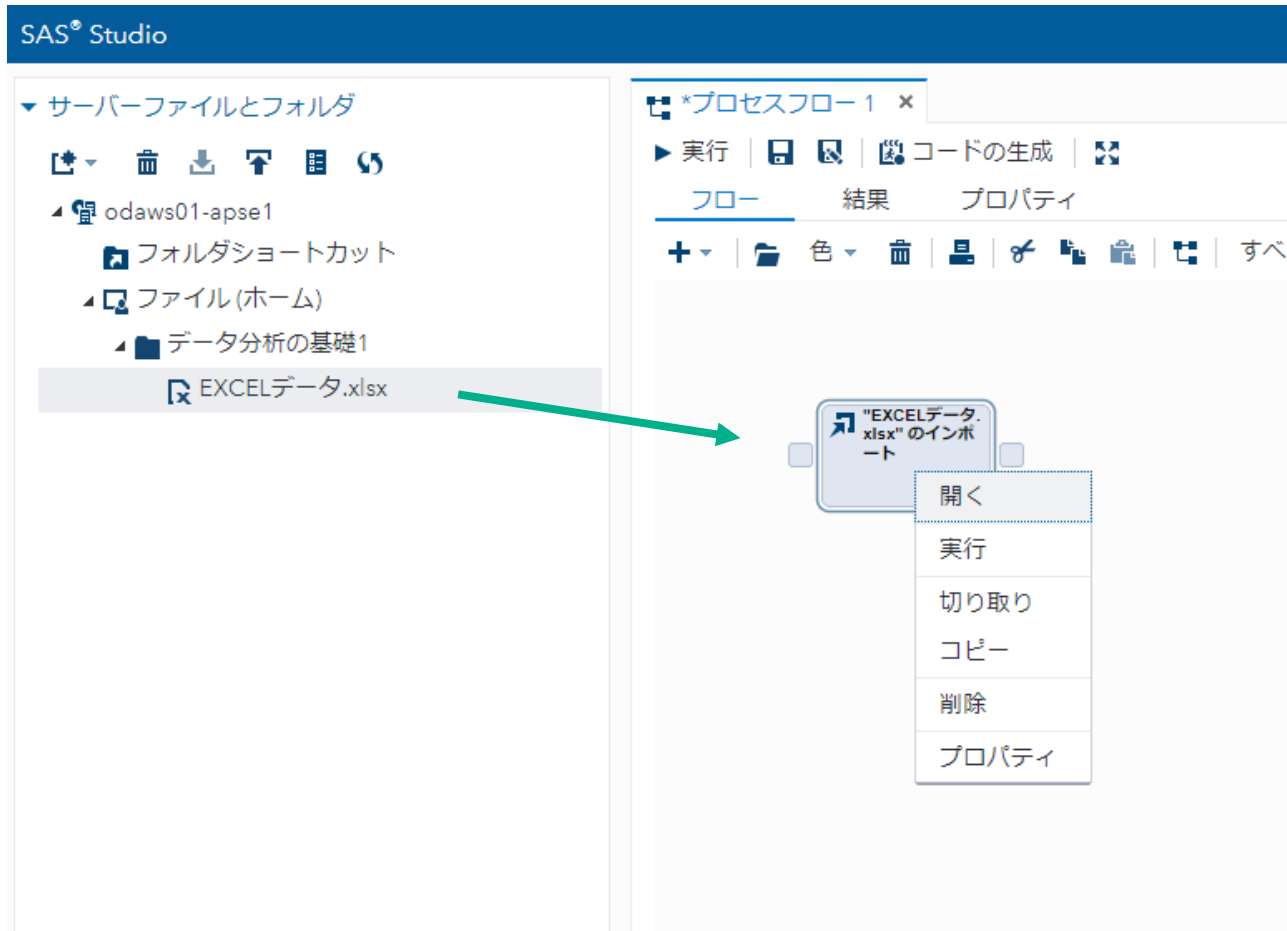


5. ファイルをアップロードする。

「データ分析の基礎1」を右クリック、「ファイルのアップロード」をクリックし、「ファイルの選択」から、ファイルを選択し「アップロード」をクリックする。



6. 「データ分析の基礎1」フォルダを開きファイル（EXCELデータ）を右側のプロセスフロー画面にドラッグし、右クリックして「開く」を選択する。



7. 「ワークシート名（Data1）」を入力し、実行ボタンをクリックする。

SAS® Studio

▼ サーバーファイルとフォルダ

- odaws01-apse1
 - フォルダショートカット
 - ファイル(ホーム)
 - データ分析の基礎1
 - EXCELデータ.xlsx

*プロセスフロー 1 ×

プロセスフロー 1 > "EXCELデータ.xlsx" のインポート

設定 | コード/結果 | 分割 | **実行** | 終了

オプション | ノード

▼ ファイル情報

ソースファイル

ファイル名: EXCELデータ.xlsx

ソースの場所: /home/u61364448/データ分析の基礎1

ワークシート名:
Data1

出力データ

SAS Server: SASApp

データセット名: IMPORT

ライブラリ: WORK

変更

▼ オプション

ファイルの種類:
デフォルト(ファイル拡張子に基づく) ▼

コード | ログ | 結果

行番号

```
1 /* 生成コード (IMPORT) */
2 /* ソースファイル: EXCELデータ.xlsx */
```

8. 「出力データ」をクリックし、「テーブル（データセット名）」、データの内容を確認する。

The screenshot shows the SAS Studio interface. On the left is the 'Server Files and Folders' pane. The main workspace is titled '*プロセスフロー 1' and shows the 'Import Excel Data' process. The 'Options' tab is active, displaying file information and output settings. The 'Output Data' tab is highlighted with a blue circle. Below the tabs, the 'Table' dropdown is set to 'WORK.IMPORT1', also highlighted with a blue circle. The 'Columns' section shows 'すべて選択' and '報酬A' selected. The 'Properties' table at the bottom is empty. The data preview table on the right shows 9 rows and 1 column.

SAS® Studio

サーバーファイルとフォルダ

- odaws01-apse1
 - フォルダショートカット
 - ファイル(ホーム)
 - データ分析の基礎1
 - EXCELデータ.xlsx

*プロセスフロー 1

プロセスフロー 1 > "EXCELデータ.xlsx" のインポート

設定 | コード/結果 | 分割

オプション | ノード

ファイル情報

ソース ファイル

ファイル名: EXCELデータ.xlsx

ソースの場所: /home/u61364448/データ分析の基礎1

ワークシート名: Data1

出力データ

SAS Server: SASApp

データセット名: IMPORT1

ライブラリ: WORK

変更

オプション

ファイルの種類: デフォルト (ファイル拡張子に基づく)

コード | ログ | 結果 | 出力データ

テーブル: WORK.IMPORT1

ビュー: 列名

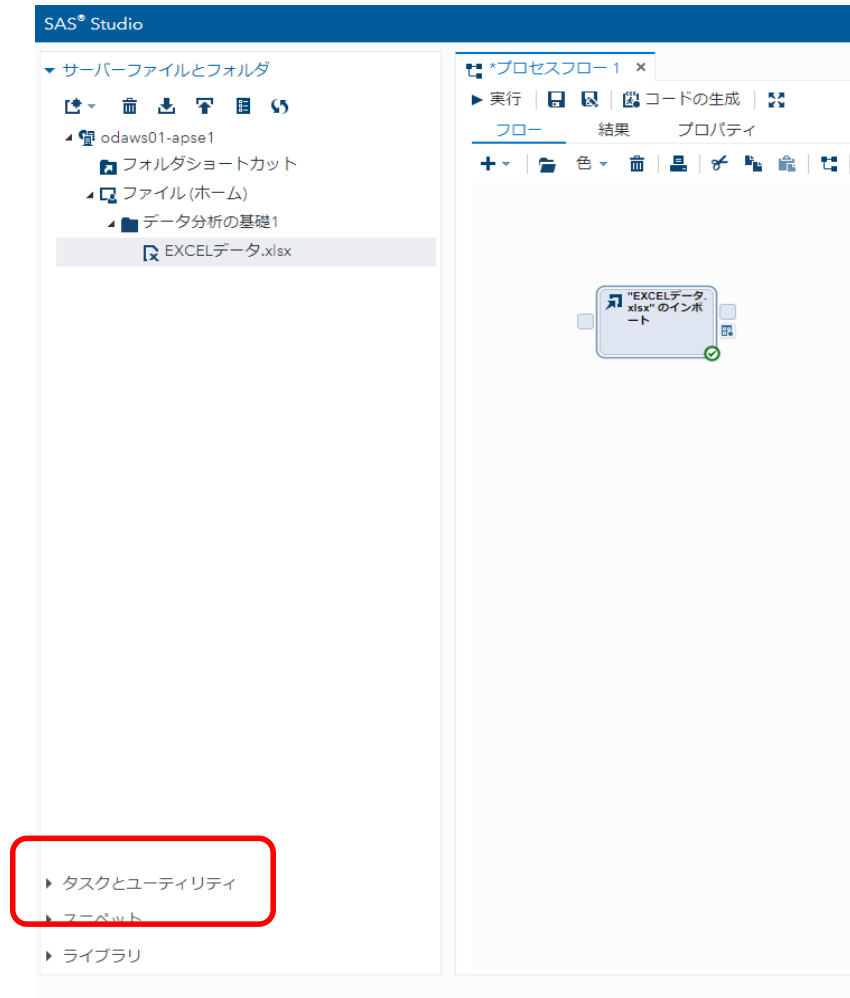
フィルタ: (なし)

合計行数: 9 合計列数: 1

プロパティ	値
ラベル	

1
2
3
4
5

9. 「プロセスフロー」をクリックしてフロー画面に戻り、「タスクとユーティリティ」を開く。



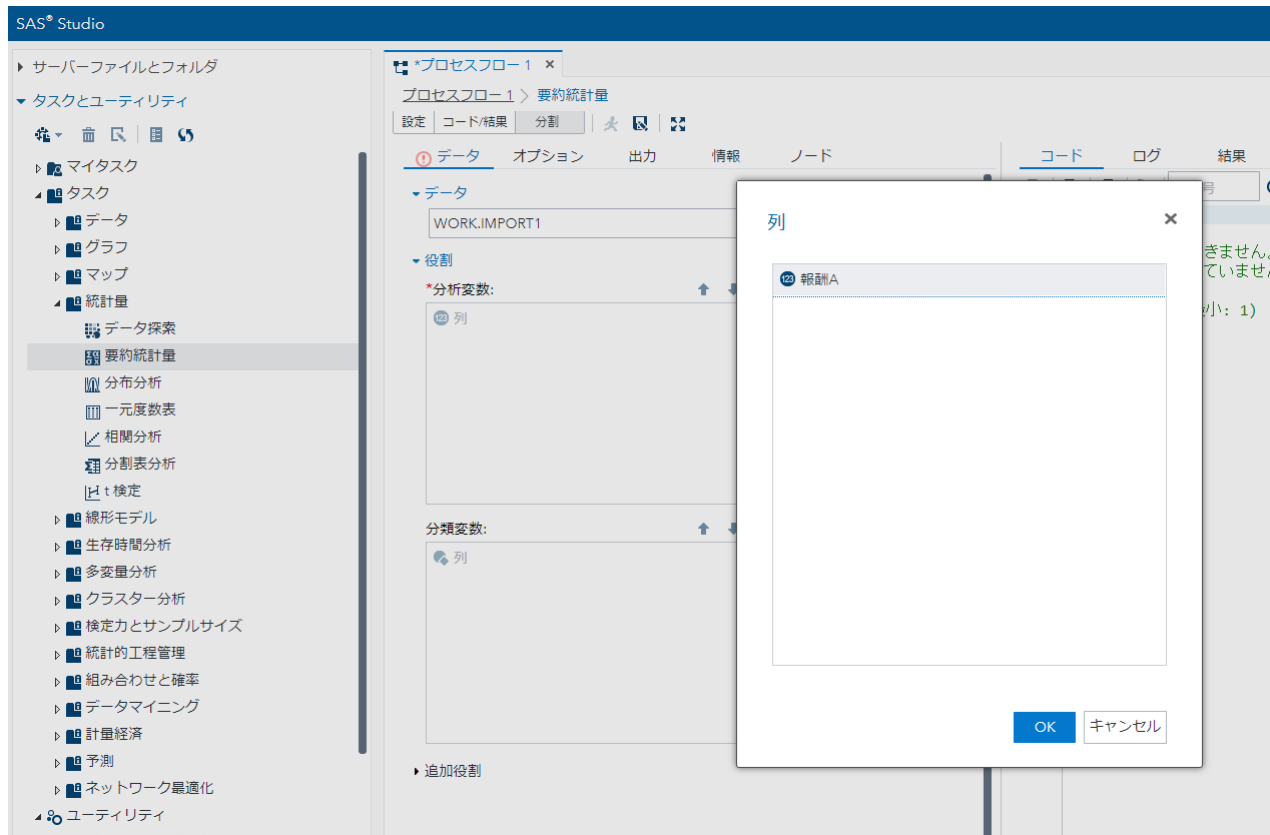
10. 「タスク」－「統計量」をクリックし、「要約統計量」をフロー画面にドラッグし「EXCELデータ・・・」と結合、右クリック－「開く」をクリックする。

The screenshot displays the SAS Studio interface. On the left, the 'タスクとユーティリティ' (Tasks and Utilities) pane is expanded to '統計量' (Statistics), where '要約統計量' (Summary Statistics) is selected. The main workspace shows a flow diagram with two tasks: 'EXCELデータ...xlsx'のインポート' (Import Excel Data...xlsx) and '要約統計量' (Summary Statistics). A dashed arrow connects the output of the first task to the input of the second. A right-click context menu is open over the '要約統計量' task, with the '開く' (Open) option highlighted. The menu also includes '実行' (Run), '切り取り' (Cut), 'コピー' (Copy), '削除' (Delete), and 'プロパティ' (Properties).

11. 「データ」の「テーブル（データセット名）」を確認する。
変更する場合は「テーブルの選択」から「データセット」を選択する。
12. 「分析変数」をセットする。（右の+マークをクリック）

The screenshot displays the SAS Studio interface. On the left, the 'Tasks and Utilities' pane shows a tree view with 'Data' selected under 'Statistics'. The main workspace shows the 'Data' tab for a process flow named '*プロセスフロー 1'. The 'Data' section displays 'WORK.IMPORT1' as the selected dataset. A red circle highlights the 'Table Selection' icon (a small table icon) next to the dataset name. A red arrow points to this icon with the text 'テーブルの選択'. Below the dataset name, the 'Analysis Variables' section is visible, showing a list of variables with a red circle highlighting the '+' button used to add variables. The 'Classification Variables' section is also visible below.

13. 分析変数を確認し、OKボタンをクリックする。



14. 「オプション」をクリックし、「パーセント点」をクリックし、出力項目にチェックを入れ、実行する。

The screenshot shows the SAS Studio interface. On the left is a navigation pane with a tree structure. The main area on the right displays the 'オプション' (Options) tab for a process flow named 'プロセスフロー 1'. The '統計量' (Statistics) section is expanded, showing '基本統計量' (Basic Statistics) and '追加統計量' (Additional Statistics). Under '追加統計量', the 'パーセント点' (Percentiles) section is expanded, listing various percentile options. A red rectangle highlights the 'パーセント点' section, and a green circle highlights the 'オプション' tab. The 'パーセント点' section includes the following options:

- ☐ 1%点
- ☐ 5%点
- ☐ 10%点
- ☒ 下側四分位点(25%点)
- ☒ 中央値(50%点)
- ☒ 上側四分位点(75%点)
- ☐ 90%点
- ☐ 95%点
- ☐ 99%点
- ☒ 四分位範囲

Below these options is a dropdown menu for '分位点の計算方法:' (Method of calculating percentiles), which is currently set to '順序統計量に基づく方法' (Method based on order statistics). The 'プロット' (Plots) and '詳細' (Details) sections are also visible at the bottom.

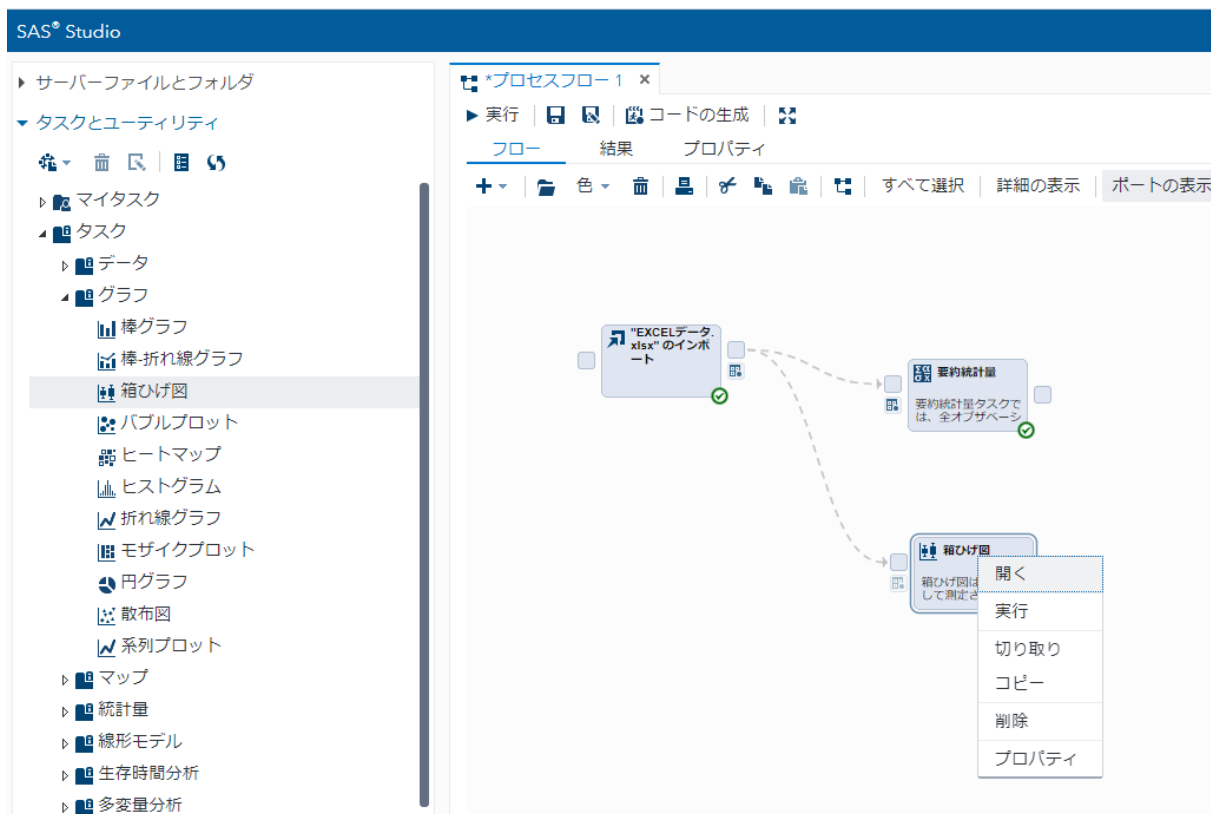
15. 要約統計量が表示される。

分析変数：報酬A 報酬A								
平均	標準偏差	最小値	最大値	中央値	N	下側四分位点	上側四分位点	四分位範囲
401.1111111	250.2056376	229.0000000	851.0000000	292.0000000	9	252.0000000	331.0000000	79.0000000

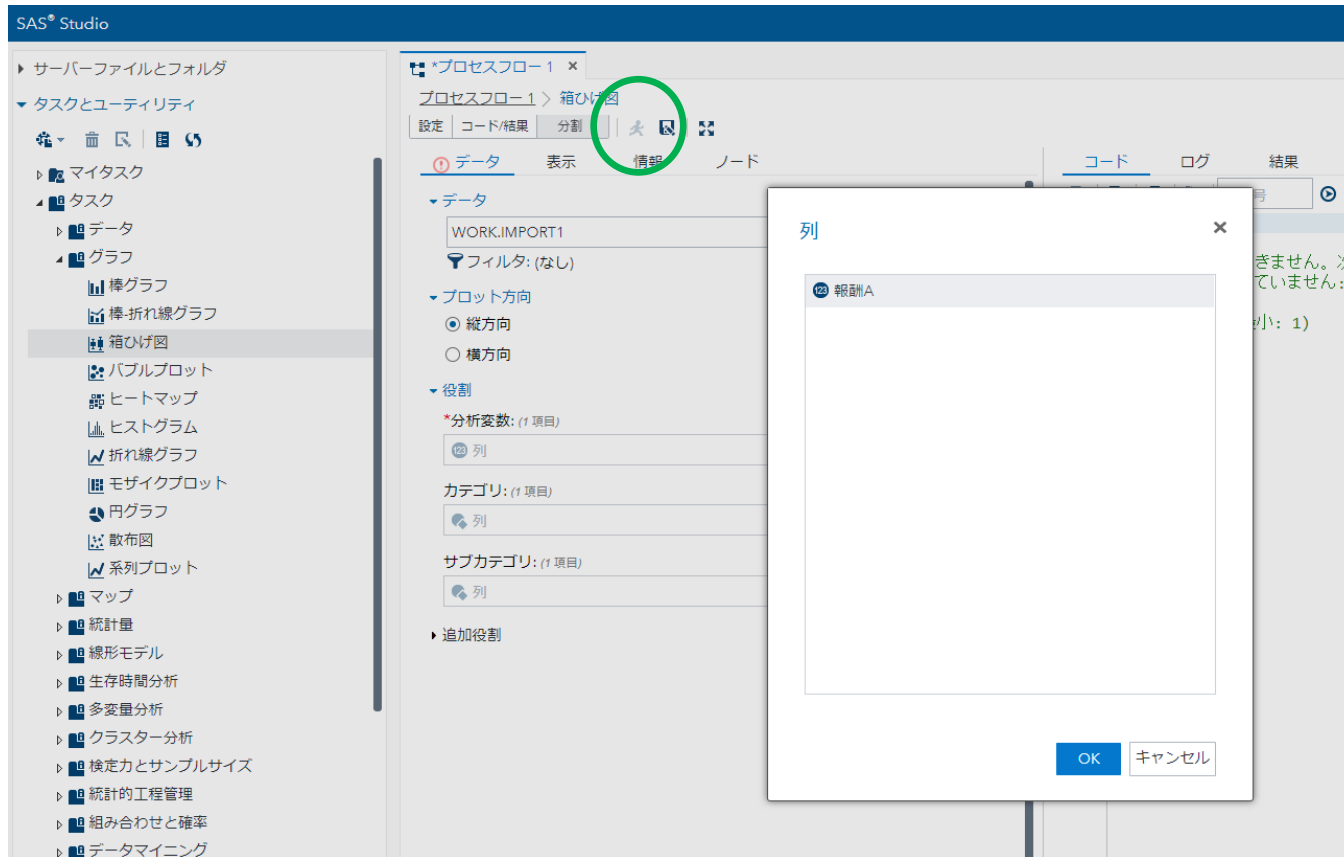
平均	401.1
Q1 四分位点 (25%)	252
Q2 中央値	292
Q3 四分位点 (75%)	331
四分位範囲	79

箱ひげ図の作成 (SAS Studio)

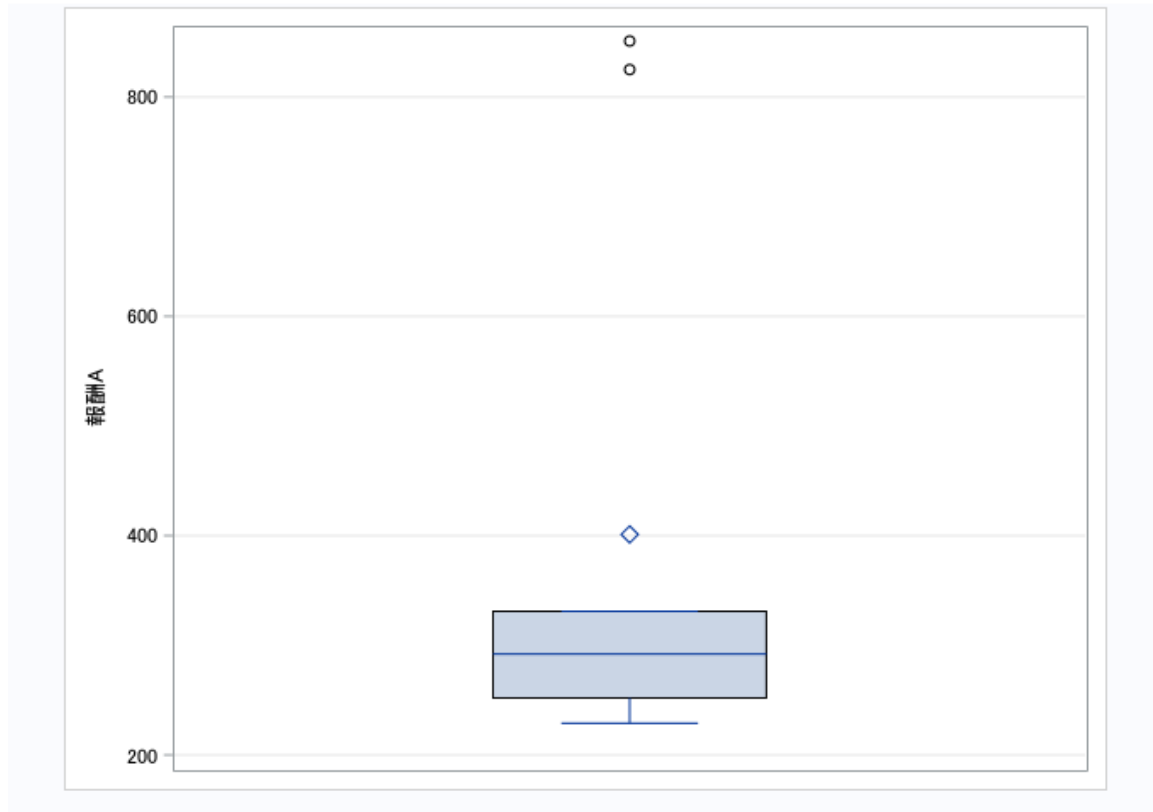
1. 「タスクとユーティリティ」→「タスク」→「グラフ」の「箱ひげ図」を選択、
2. フロー画面にドラッグし、「EXCELデータ・・・」と結合、右クリック「開く」をクリックする。



3. 「データ」の「テーブル（データセット名）」を確認する。
変更する場合は「テーブルの選択」から「データセット」を選択する。
4. 「分析変数」を確認し、実行ボタンをクリックする。



5. 箱ひげ図が表示される。



<事例>

1 2 3 4 5 6 7 8 9 10 11 12 13

3	11	11	12	13	13	14	20	25	28	50	75	92
---	----	----	----	----	----	----	----	----	----	----	----	----

Q1 四分位点 (25%) : 12

Q2 中央値 : 14

Q3 四分位点 (75%) : 28

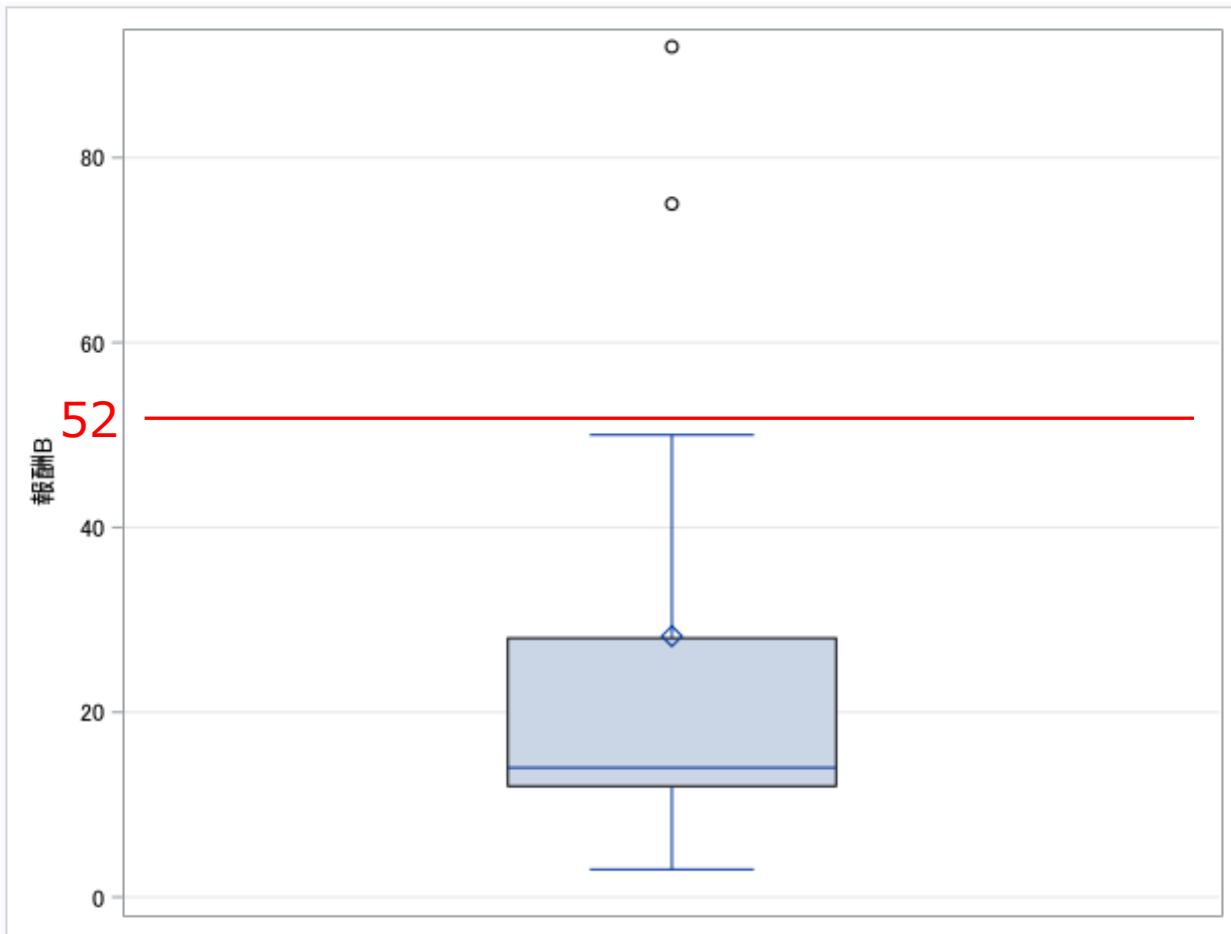
四分位範囲 $Q3 - Q1 = 28 - 12 = 16$

$12 - 1.5 \times 16 = -12$ $28 + 1.5 \times 16 = 52$

外れ値 : -12以下 52以上

分析変数：報酬B 報酬B								
平均	標準偏差	最小値	最大値	中央値	N	下側四分位点	上側四分位点	四分位範囲
28.2307692	27.3348186	3.0000000	92.0000000	14.0000000	13	12.0000000	28.0000000	16.0000000

平均	28.2
Q1 四分位点 (25%)	12
Q2 中央値	14
Q3 四分位点 (75%)	28
四分位範囲	16



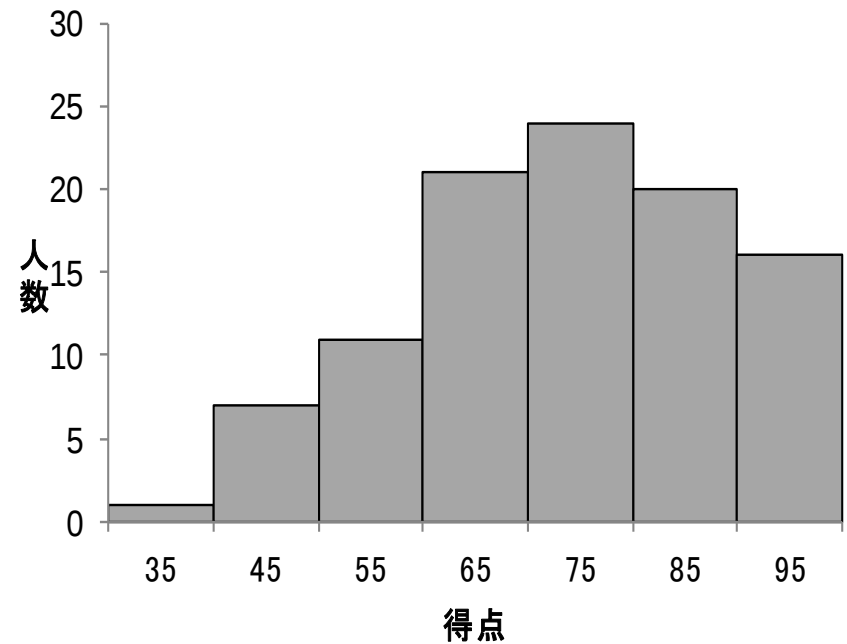
ヒストグラム

成績

88	88	60	86	86
48	86	60	46	40
64	84	70	34	62
78	36	77	72	76
65	68	90	90	56
72	54	54	52	66
68	68	60	74	72
78	56	62	96	90
36	86	48	48	90
48	68	84	66	46

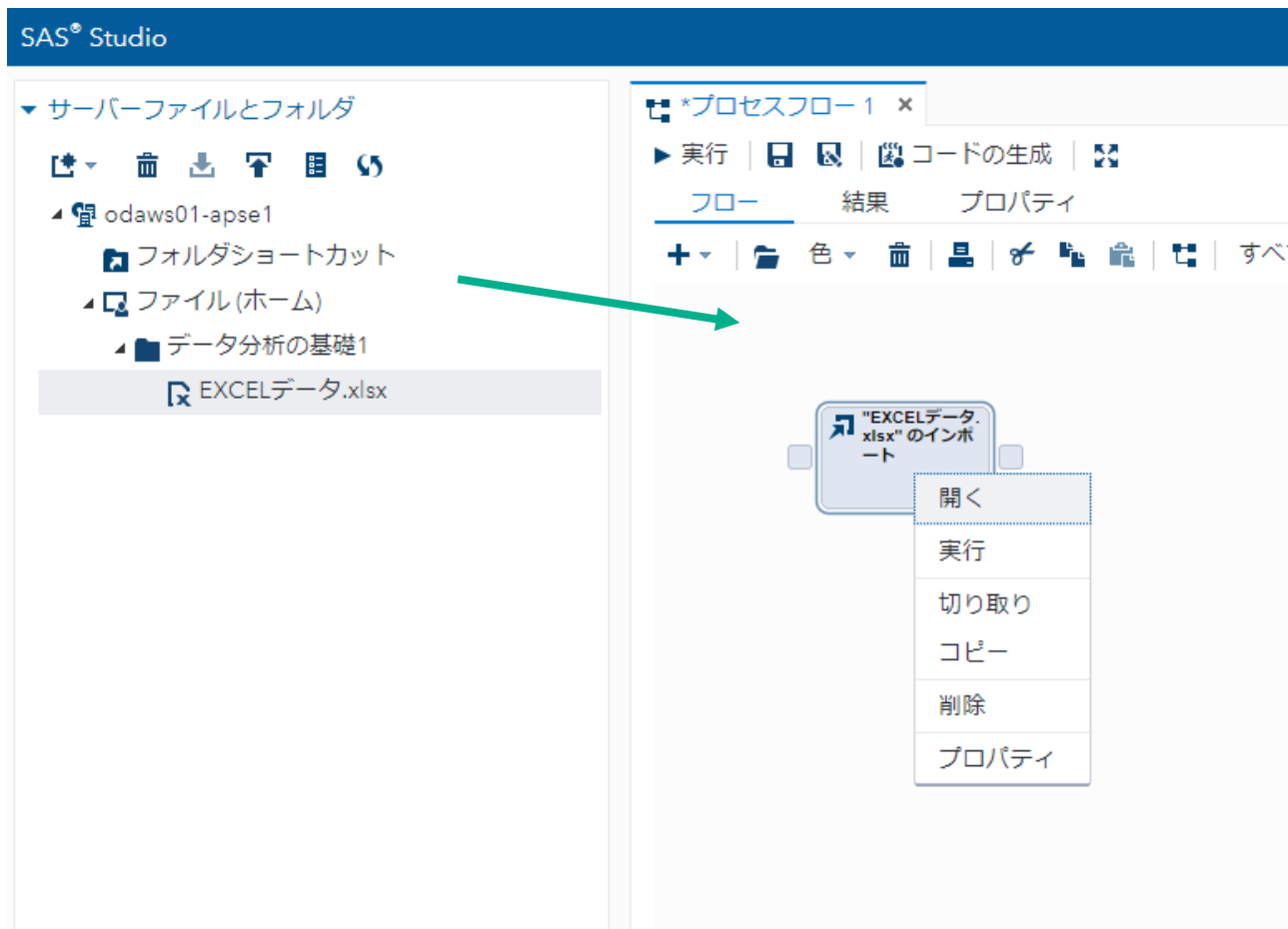
度数分布表

下限		上限	階級値	度数
30	～	39	35	1
40	～	49	45	7
50	～	59	55	11
60	～	69	65	21
70	～	79	75	24
80	～	89	85	20
90	～	99	95	16



ヒストグラムの作成 (SAS Studio)

1. 「データ分析の基礎1」フォルダを開きファイル（EXCELデータ）を右側のプロセスフロー画面にドラッグし、右クリックして「開く」を選択する。



2. 「ワークシート名（Data5）」を入力し、実行ボタンをクリックする。



3. 「出力データ」をクリックし、「テーブル（データセット名）」、データの内容を確認する。

SAS® Studio

サーバーファイルとフォルダ

- odaws01-apse1
 - フォルダショートカット
 - ファイル (ホーム)
 - データ分析の基礎1
 - EXCELデータ.xlsx

*プロセスフロー 1 x

プロセスフロー 1 > "EXCELデータ.xlsx" のインポート

設定 | コード/結果 | 分割 | 実行 | 出力 | 出力データ

オプション | ノード

ファイル情報

ソース ファイル

ファイル名: EXCELデータ.xlsx

ソースの場所: /home/u61364448/データ分析の基礎1

ワークシート名: Data5

出力データ

SAS Server: SASApp

データセット名: IMPORT1

ライブラリ: WORK

変更

オプション

ファイルの種類: デフォルト (ファイル拡張子に基づく)

コード | ログ | 結果 | 出力データ

テーブル: WORK.IMPORT1 | ビュー: 列名 | フィルタ: (なし)

合計行数: 50 合計列数: 1

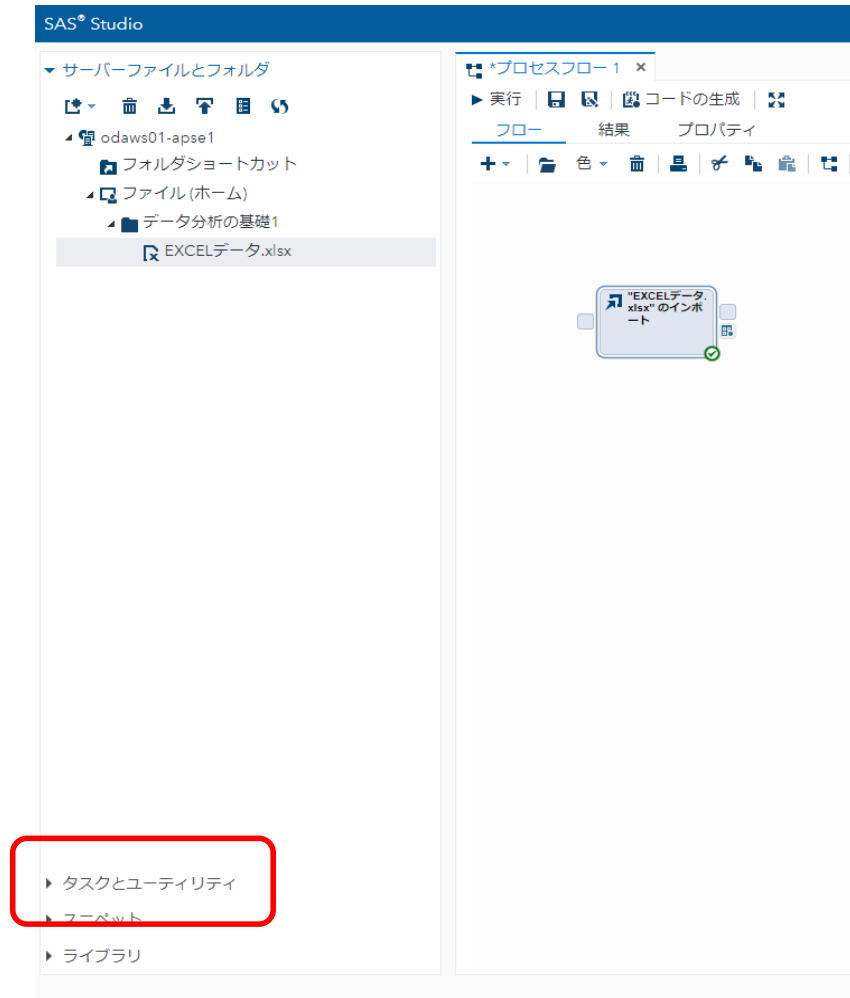
すべて選択

成績

プロパティ | 値

成績
1
2
3
4

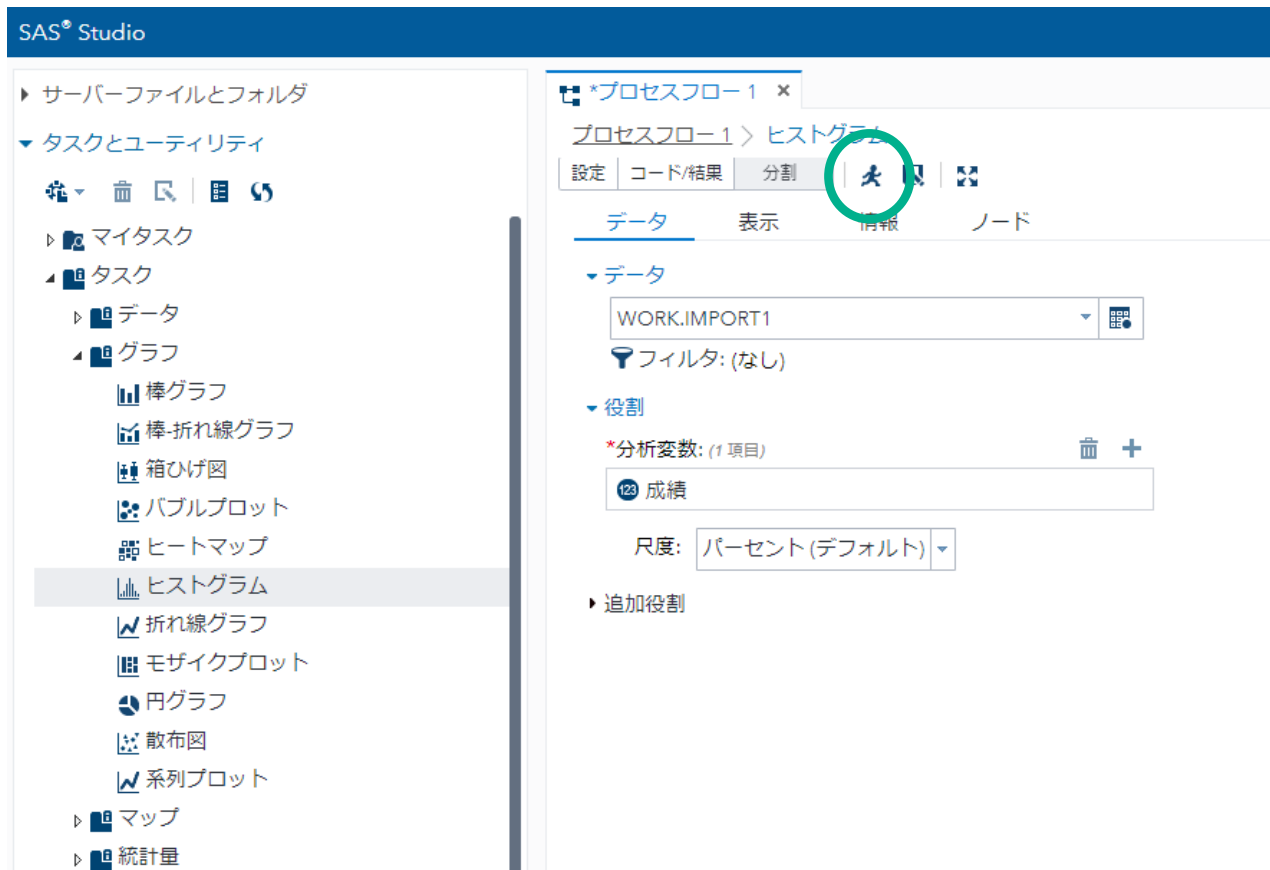
4. 「プロセスフロー」をクリックしてフロー画面に戻り、「タスクとユーティリティ」を開く。



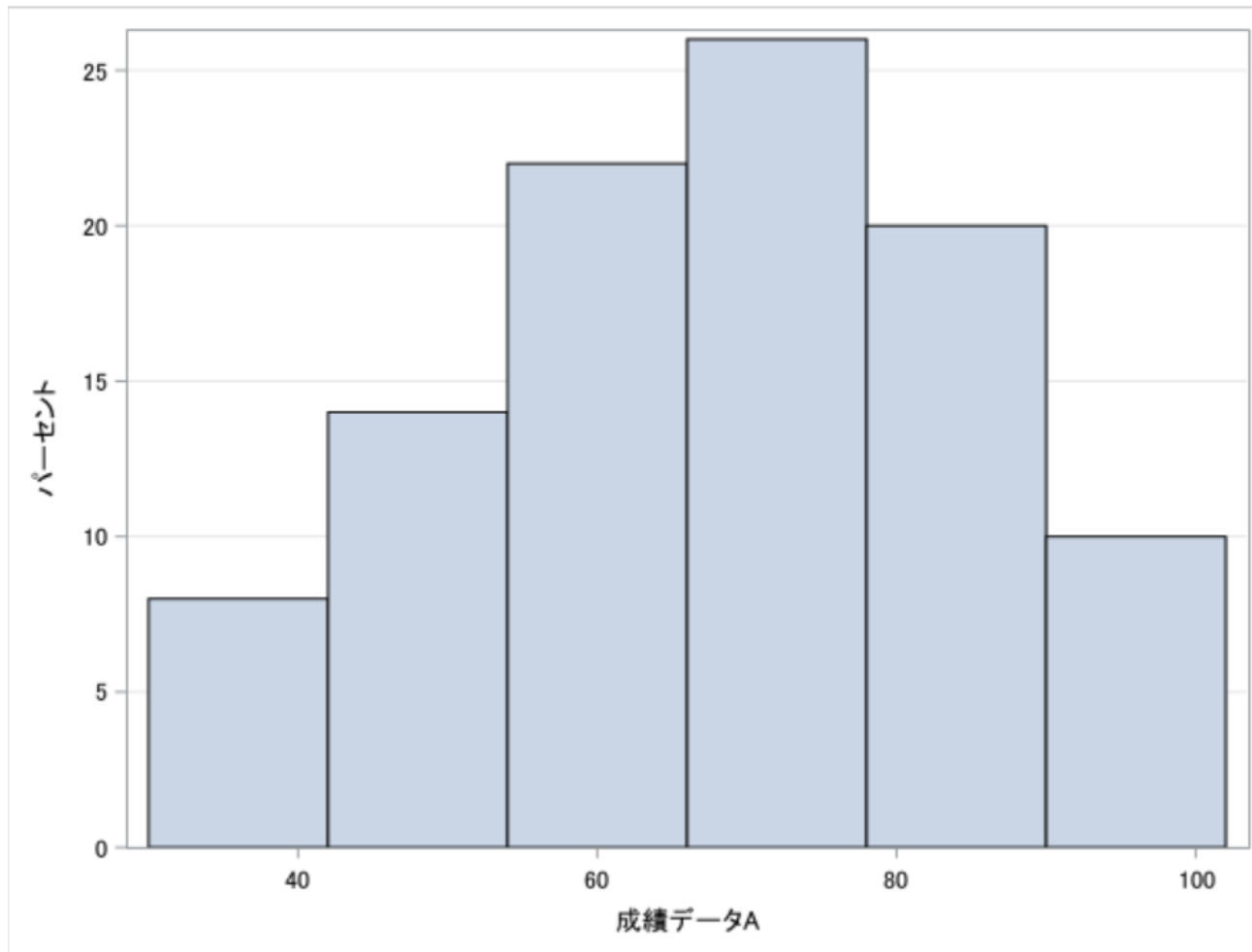
5. 「タスク」－「グラフ」をクリックし、「ヒストグラム」をフロー画面にドラッグし「EXCELデータ・・・」と結合、右クリック－「開く」をクリックする。

The screenshot shows the SAS Studio interface. On the left is the 'タスクとユーティリティ' (Tasks and Utilities) pane, which is expanded to show 'グラフ' (Graphs). Under 'グラフ', 'ヒストグラム' (Histogram) is selected and highlighted. On the right is the main workspace titled '*プロセスフロー 1' (Process Flow 1). It contains a flow diagram with two tasks: 'EXCELデータ.xlsxのインポート' (Import Excel data.xlsx) and 'ヒストグラム' (Histogram). A dashed arrow connects the output of the first task to the input of the second task. A right-click context menu is open over the 'ヒストグラム' task, with the '開く' (Open) option selected. The menu also includes '実行' (Run), '切り取り' (Cut), 'コピー' (Copy), '削除' (Delete), and 'プロパティ' (Properties).

6. 「データ」の「テーブル（データセット名）」を確認する。
変更する場合は「テーブルの選択」から「データセット」を選択する。
7. 「分析変数」をセットし、実行ボタンをクリックする。



8. ヒストグラムが表示される。



分析の基本的な考え方

英語の研修前後の成績（20人）

	A君の得点	全体の平均点	得点-平均点
研修前	70	58.3	+11.7
研修後	75	58.3	+16.7

A君の成績の評価は？

平均を評価基準とする合理性？

全員の成績

- <研修前>

70、56、89、27、69、57、69
50、33、67、37、49、98、69
68、25、65、67、33、68

- <研修後>

75、31、95、36、89、88、89
76、28、47、23、28、96、48
51、20、30、91、27、98

研修後のバラツキが大きい！

全員の成績

- ・＜研修前＞

70、56、89、27、69、57、69

50、33、67、37、49、98、69

68、25、65、67、33、68

- ・＜研修後＞

75、31、95、36、89、88、89

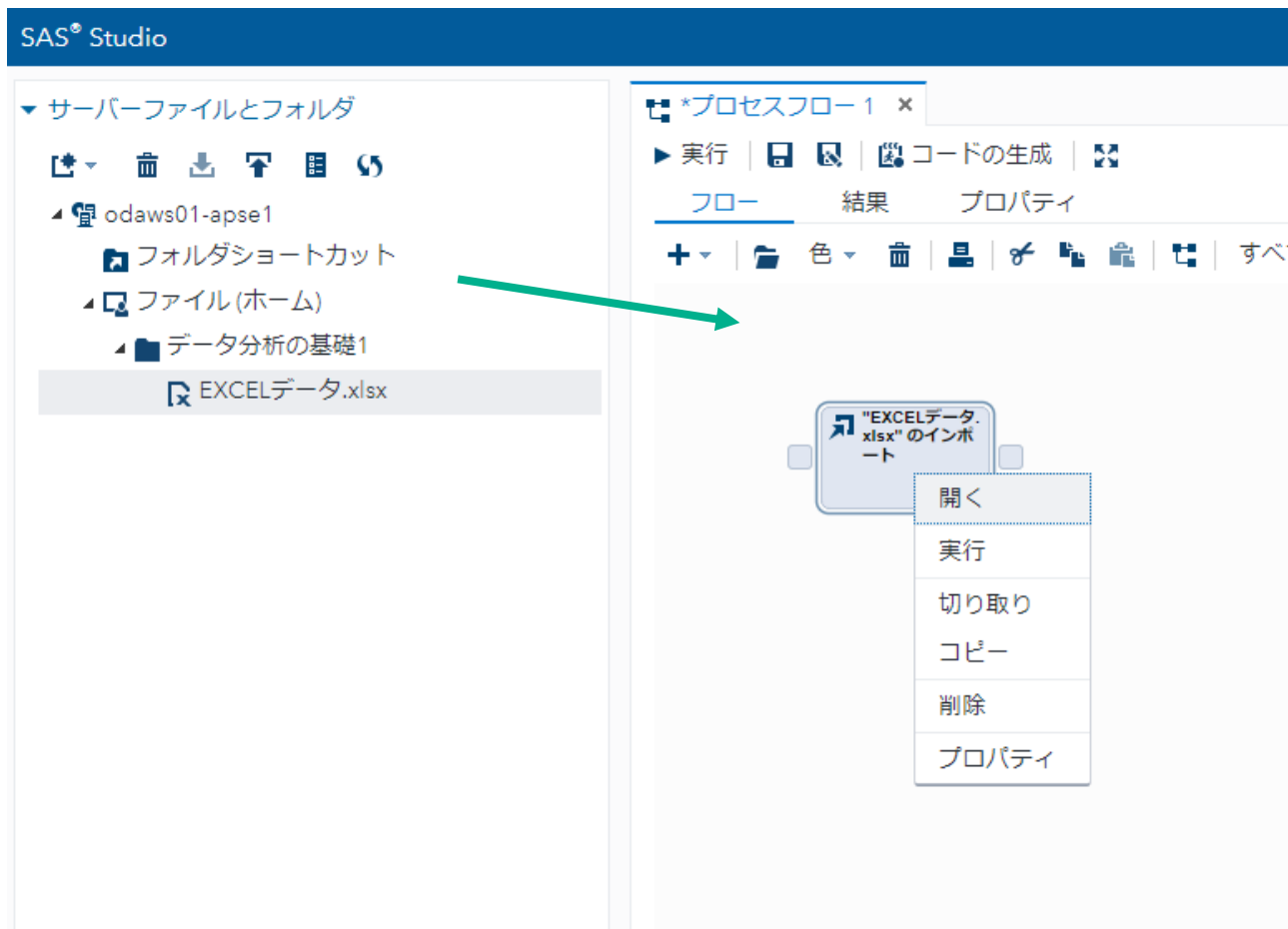
76、28、47、23、28、96、48

51、20、30、91、27、98

研修後の成績順位は下がった！

要約統計量 (SAS Studio)

1. 「データ分析の基礎1」フォルダを開きファイル (EXCELデータ) を右側のプロセスフロー画面にドラッグし、右クリックして「開く」を選択する。



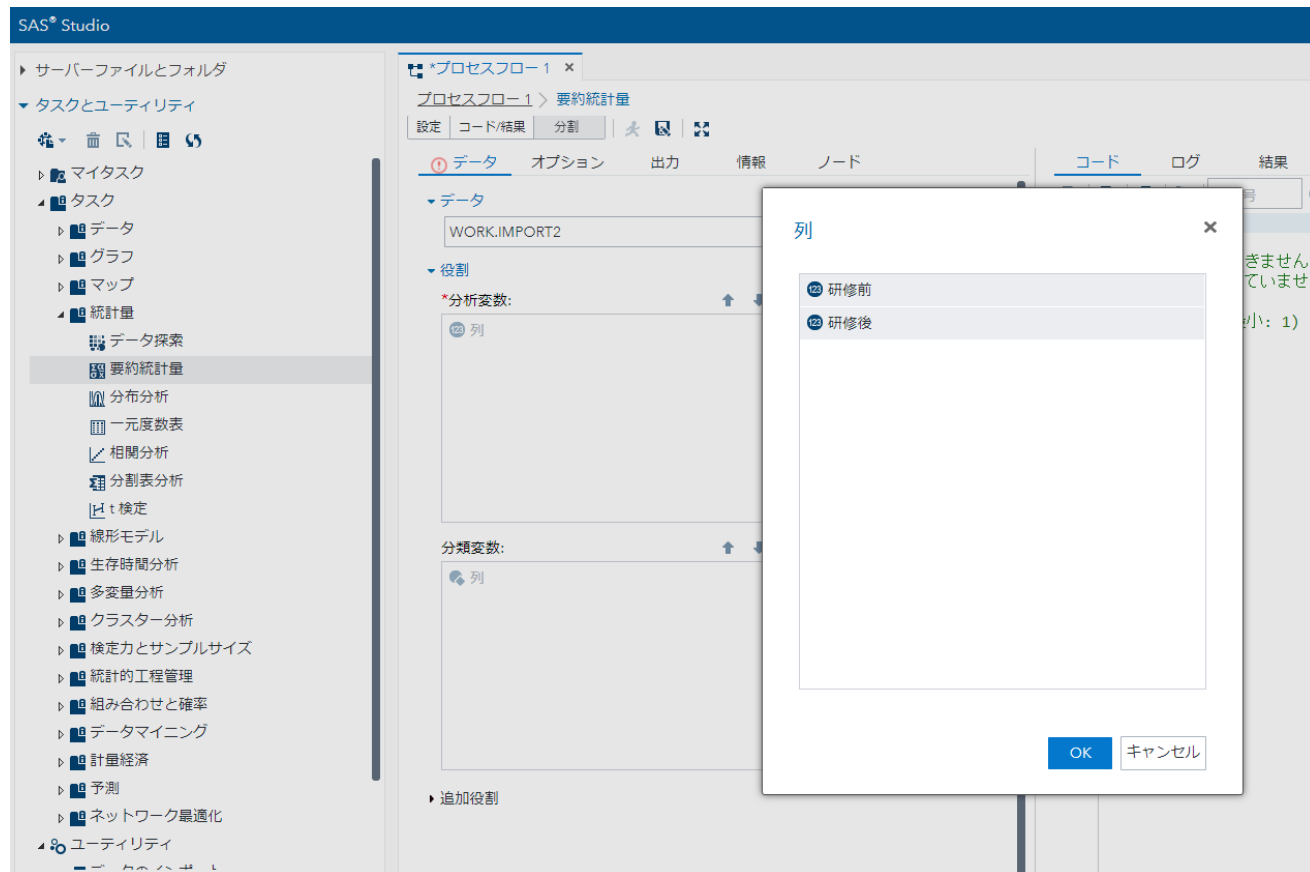
2. 「ワークシート名（Data8）」を入力し、実行ボタンをクリックする。



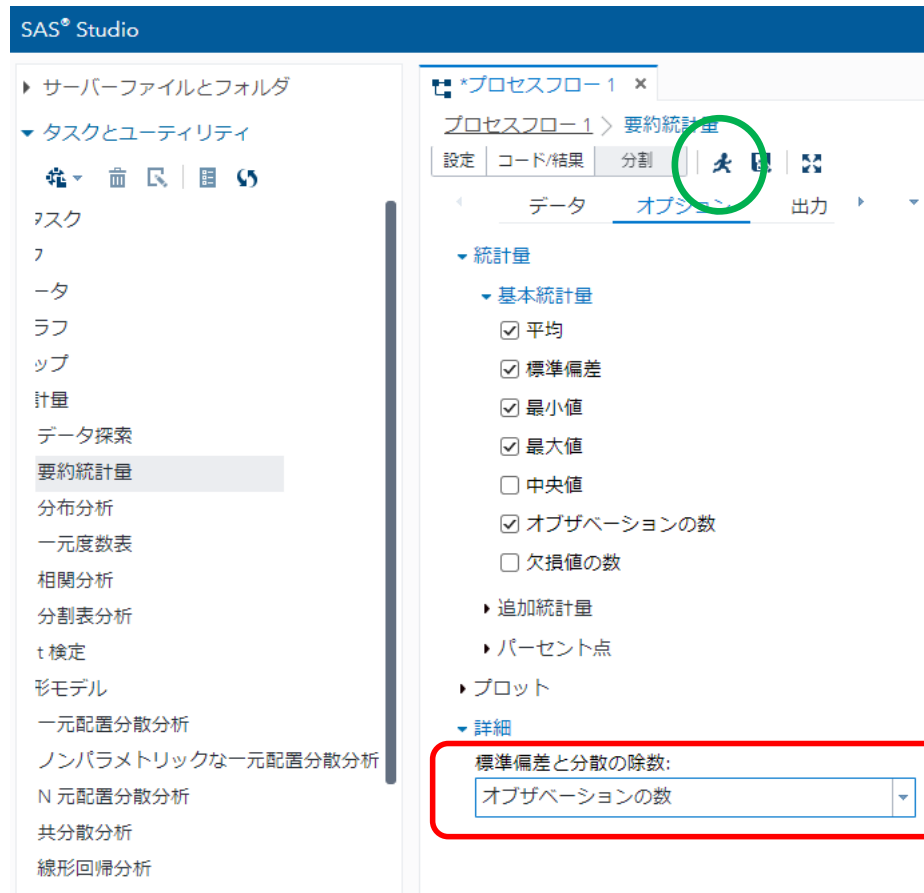
3. 「タスク」－「統計量」をクリックし、「要約統計量」をフロー画面にドラッグし「EXCELデータ・・・」と結合、右クリック－「開く」をクリックする。

The screenshot displays the SAS Studio interface. On the left, the 'タスクとユーティリティ' (Tasks and Utilities) pane is expanded to '統計量' (Statistics), with '要約統計量' (Summary Statistics) selected. The main workspace shows a flow diagram with two tasks: 'EXCELデータ...xlsx'のインポート' (Import EXCEL data...xlsx) and '要約統計量' (Summary Statistics). A dashed arrow connects the output of the first task to the input of the second. A right-click context menu is open over the '要約統計量' task, showing options: '開く' (Open), '実行' (Run), '切り取り' (Cut), 'コピー' (Copy), '削除' (Delete), and 'プロパティ' (Properties).

4. 「データ」の「テーブル（データセット名）」を確認する。
変更する場合は「テーブルの選択」から「データセット」を選択する。
5. 「分析変数」をセットする。



6. 「オプション」をクリックし、「統計量」→「基本統計量」の出力項目にチェックを入れ選択する。
「詳細」をクリックし、「標準偏差と分散の除数」から「オブザベーションの数」を選択、実行する。



7. 要約統計量が表示される。

変数	ラベル	平均	標準偏差	最小値	最大値	N
研修前	研修前	58.3000000	19.1940095	25.0000000	98.0000000	20
研修後	研修後	58.3000000	28.8376490	20.0000000	98.0000000	20

<標準偏差>

研修前 : 19.19

研修後 : 28.84

A君の成績の評価

	成績	平均	成績－平均	標準偏差
研修前	70	58.3	11.7	19.19
研修後	75	58.3	13.7	28.84

研修前

$$\frac{70-58.3}{19.19} = 0.610 >$$

研修後

$$\frac{75-58.3}{28.84} = 0.579$$

成績と平均値の差

標準偏差



Z値

Z 値の比較によりデータの評価・比較が可能

$$Z \text{ 値} \times 10 + 50$$



偏差値

$$\text{研修前の偏差値} = 0.610 \times 10 + 50 = 56.09$$

$$\text{研修後の偏差値} = 0.579 \times 10 + 50 = 55.79$$

Z 值 $\xrightarrow{(-)}$ 0 $\xrightarrow{(+)}$

偏差值 _____ 50 _____

Z 値は、平均値 = 0

平均値より大のときプラスの値、小のときマイナスの値

偏差値は、平均値 = 50

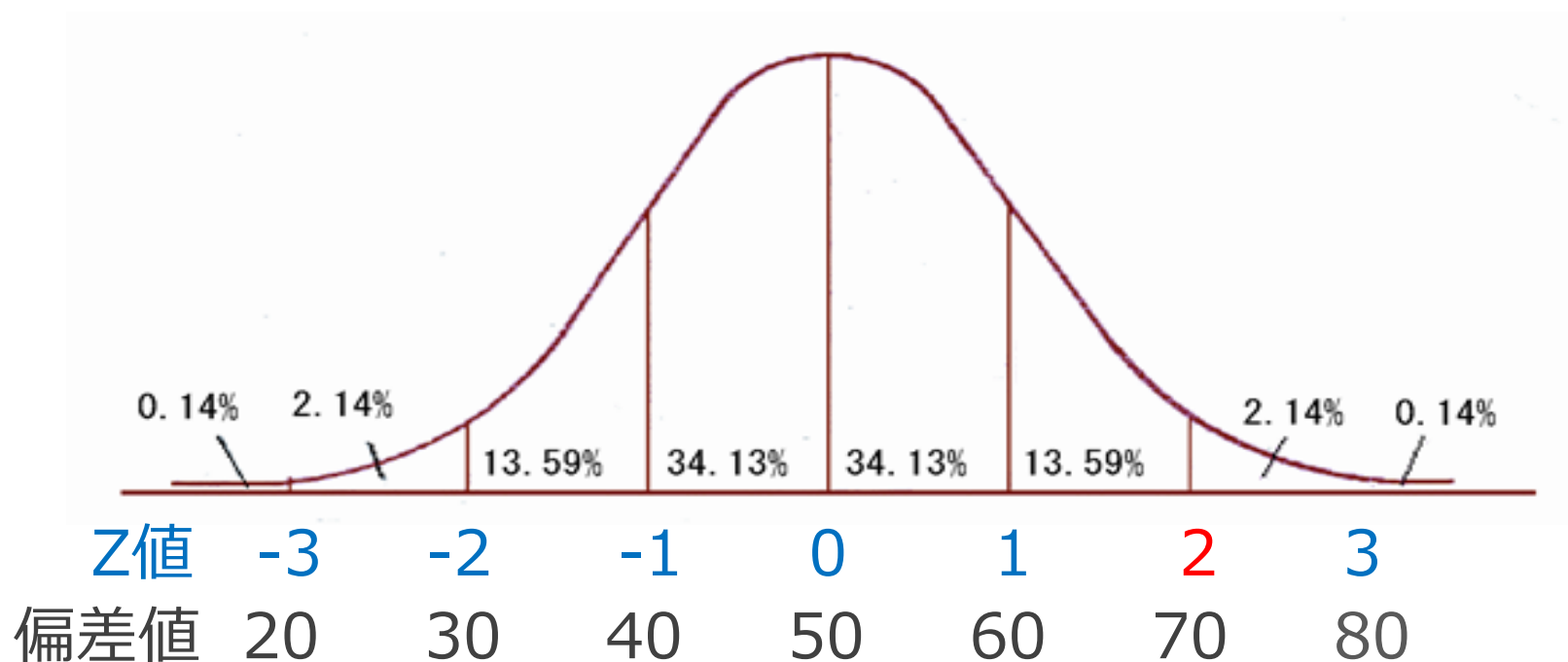
平均値より大のとき50より大きな値、平均値より小のとき50より小さな値

Z 値の特性

データ			Z 値	データ			Z 値
	1		-1.4142		3		-1.1844
	2		-0.7071		8		0.55739
	3		0		6		-0.1393
	4		0.70711		4		-0.8361
	5		1.41421		11		1.60249
平均値	3		0	6.4			0
標準偏差	1.41421		1	2.87054			1

Z 値の平均値 = 0 標準偏差 = 1

標準正規分布 (standard normal distribution)

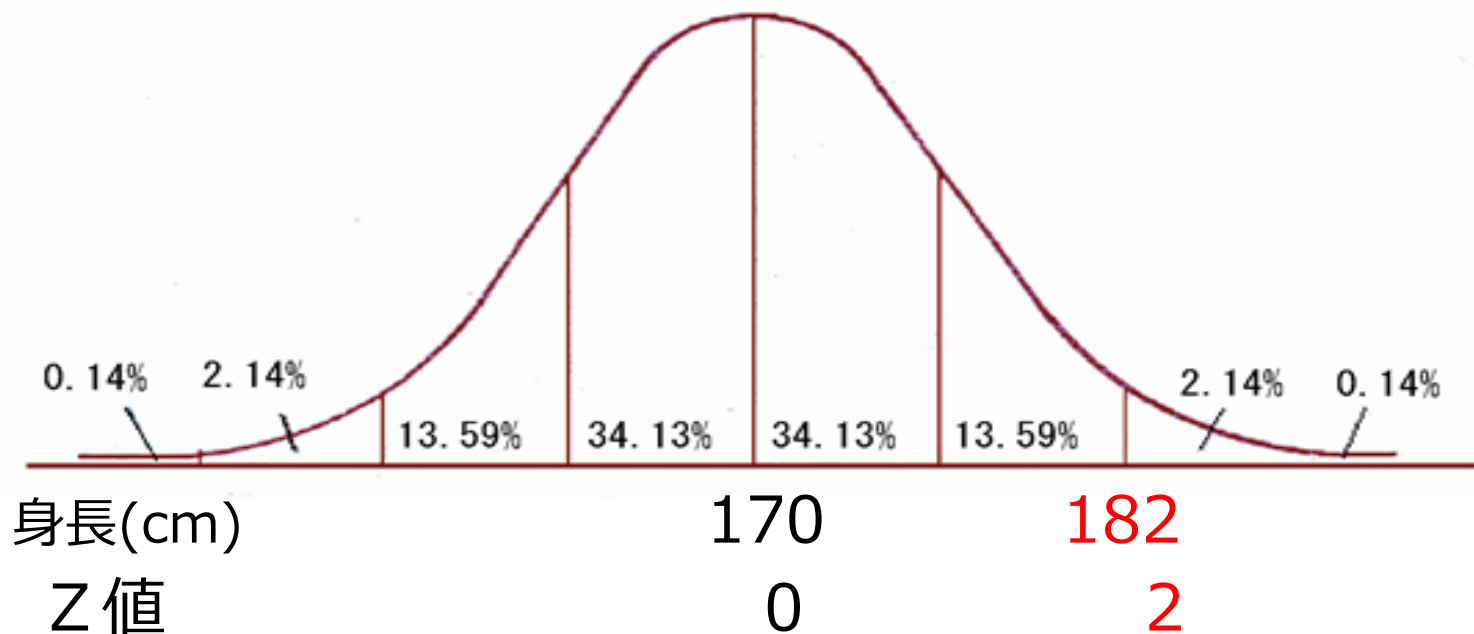


Z 値 = 1 (偏差値60) 以上の割合は約15.87%

Z 値 = 2 (偏差値70) 以上の割合は全体の約2.28%

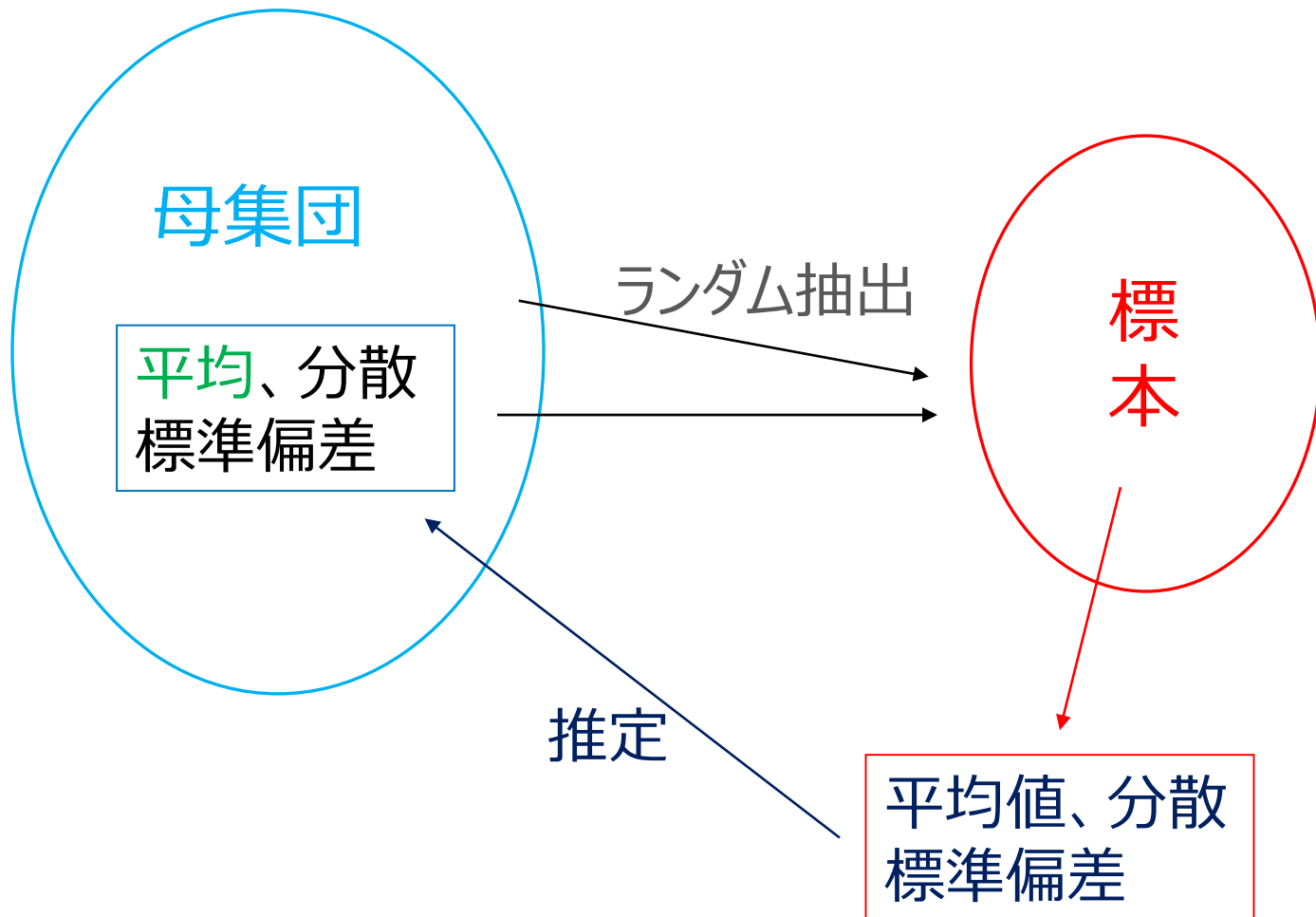
標準正規分布の活用

- 日本の成年男子の身長が平均170cm 標準偏差6cm、
182cm以上の人は、全体の何%を占めるか？



$$Z = \frac{182 - 170 \text{ (cm)}}{6 \text{ (cm)}} = \mathbf{2} \quad 182\text{cm以上の人は約}2.28\%$$

母集団と標本

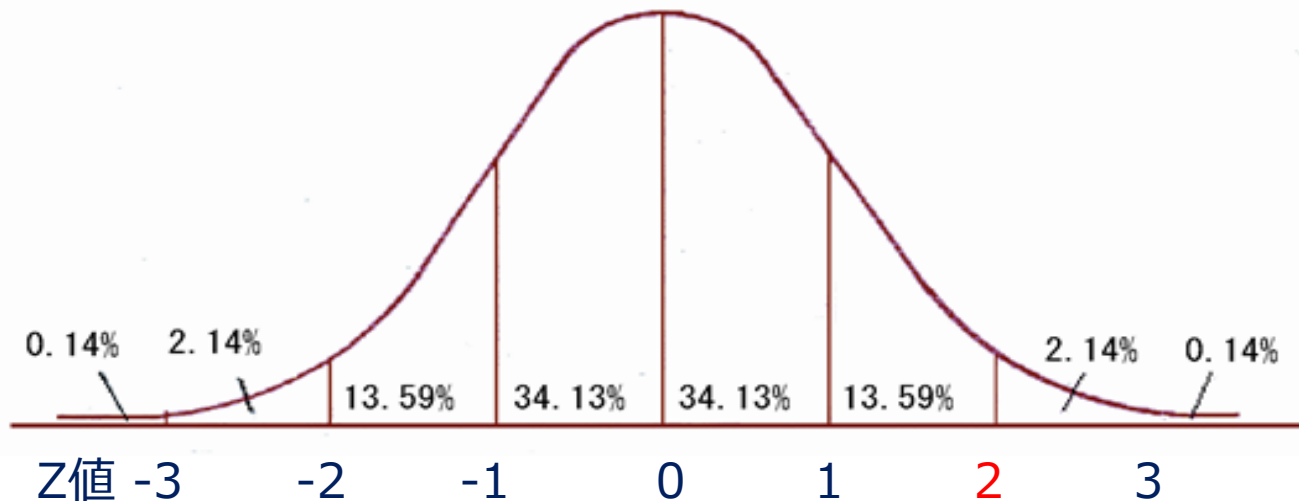


◇競合メーカー 有効成分含有量
平均 = 60mg、標準偏差 = 5mg

市場から4つ購入した平均 : 65mg
60mgより多いのか？



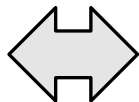
P値（有意確率）を求める！
違うと判定したときの危険率



$$Z = \frac{65-60}{5} \times \sqrt{4} = 2$$

P値（有意確率） = 2.28%

サンプルサイズが大きくなると、Z 値も大きくなる

Z 値による検定  正規分布

Z 値による検定は標本数が少ないときに精度が
良くない。

t 値による検定



t 分布

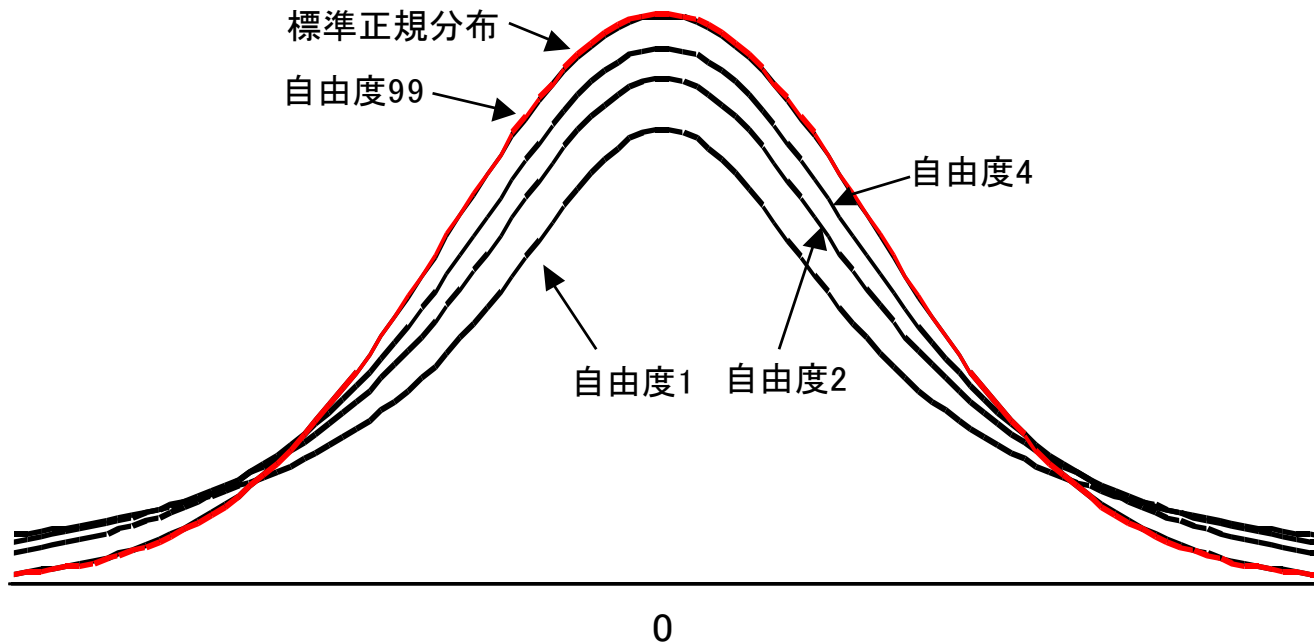
t 分布とは

1908年、W.S.ゴセット（ギネスビールの研究員）が考案。
データが少ないときに精度を上げるために t 値を使う。



- 全て（データの大小に関係なく） t 分布を使う。
- この検定方法を t 検定という。

t 分布



データの大きさによって形が変わる。
大きくなると正規分布と同じ形。

注) 自由度 : データ数 - 1

$$t \text{ 値} = \frac{\textcircled{1} \text{差 (65-60)}}{\textcircled{2} \text{バラツキ (5)}} \times \textcircled{3} \sqrt{n \text{ (4)}}$$

- ① 差が大 $\Rightarrow$ t 値は大
- ② バラツキが小 $\Rightarrow$ t 値は大
- ③ n (サンプルサイズ) が大 $\Rightarrow$ t 値は大

t 値が大 = P値 (有意確率) は小

t 値が小 = P値 (有意確率) は大

1標本 t 検定

(A) ランダムに選んだ製品10個の有効成分含有量
60mg（母平均）と違いはみられるか？

65	62	57	61	60	58	63	67	62	63
----	----	----	----	----	----	----	----	----	----

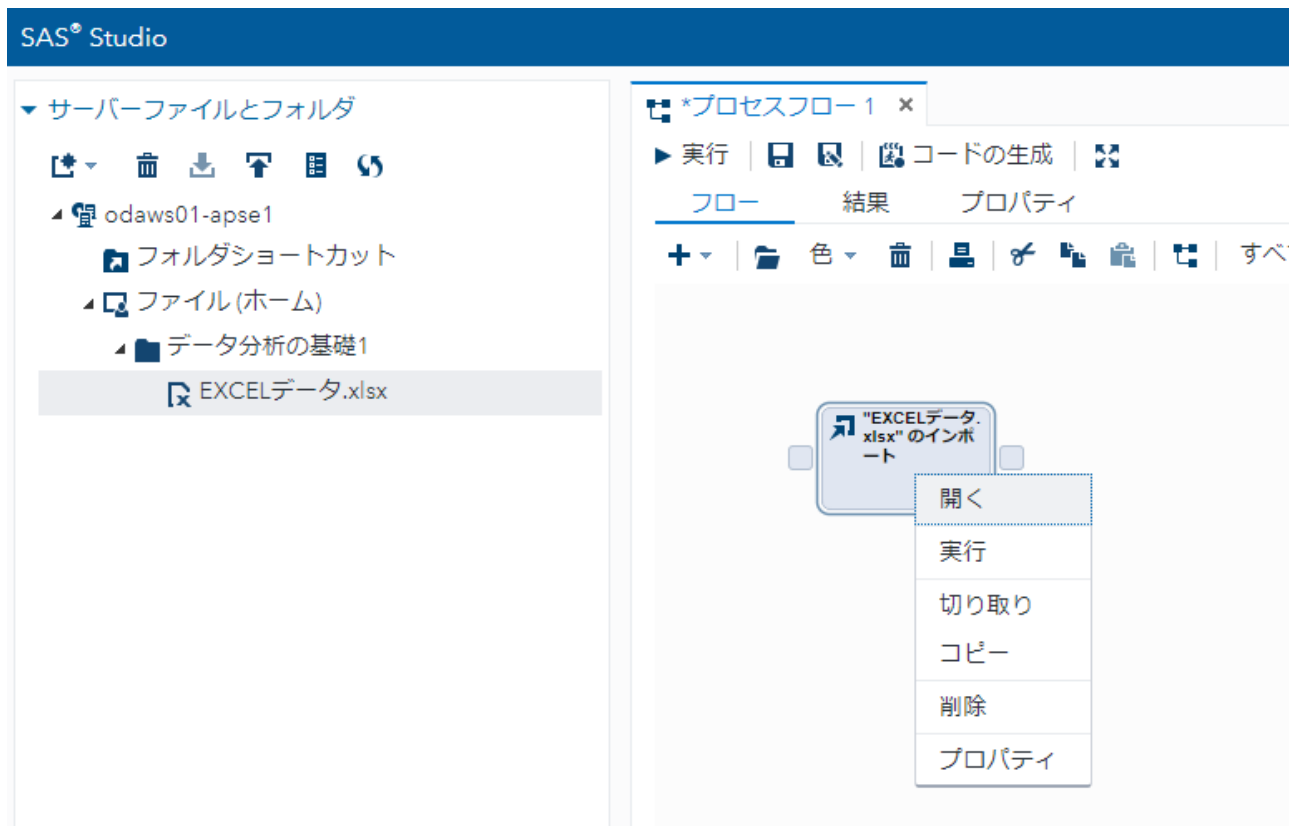
10個の平均値を求めると

$$(65+62+\cdots+62+63) \div 10 = 61.8$$

標本平均（61.8）と母平均（60.0）の比較

1標本 t 検定 (SAS Studio)

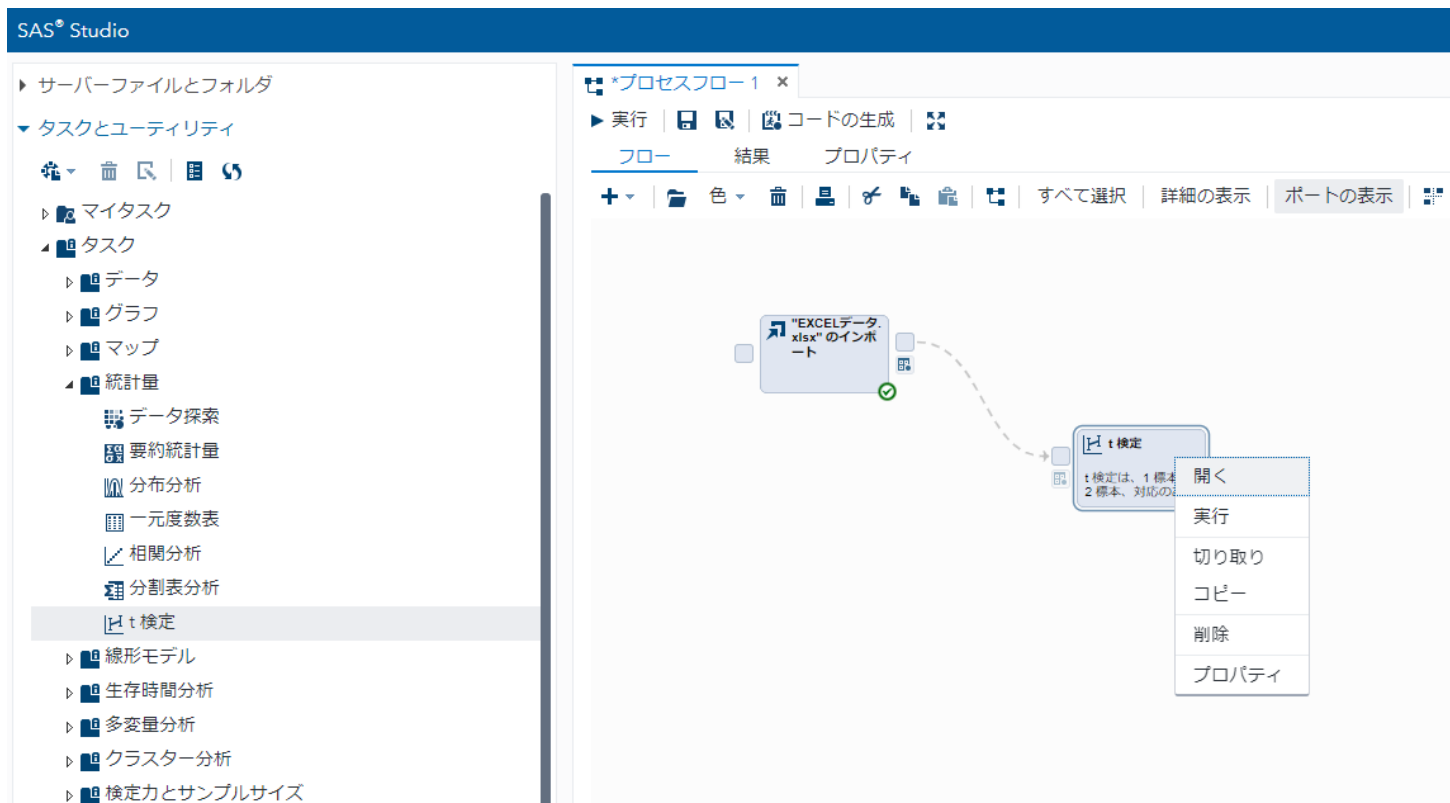
1.「データ分析の基礎1」フォルダを開きファイル（EXCELデータ）を右側のプロセスフロー画面にドラッグし、右クリックして「開く」を選択する。



2. 「ワークシート名（Data10）」を入力し、実行ボタンをクリックする。



3. 「タスクとユーティリティ」→「タスク」→「統計量」の「t 検定」をフォロー画面にドラッグし、「EXCELデータ・・・」と結合、右クリック「開く」をクリックする。



4. 「t検定」を右クリック、開き、「データ」、「t 検定」、「分析変数」をセットする。

The screenshot displays the SAS Studio interface. On the left, the 'タスクとユーティリティ' (Tasks and Utilities) pane is expanded, showing the 'タスク' (Tasks) category. Under 'タスク', the '統計量' (Statistics) sub-category is expanded, and 't検定' (t-test) is selected. The main workspace on the right shows the configuration for the 't検定' task. The 'データ' (Data) tab is active, displaying 'WORK.IMPORT' as the data source and 'フィルタ: (なし)' (Filter: (none)). The '役割' (Role) tab is also visible, showing 't検定: 1 標本検定' (t-test: 1 sample test). Under the '分析変数: (1 項目)' (Analysis variable: (1 item)) section, the variable '含有量' (Content) is listed.

5. 「オプション」をクリックし、 $\mu = 60$ を入力し、実行ボタンをクリックする。

*プロセスフロー 1 × | H *t 検定 ×

設定 | コード/結果 | 分割 | **実行** | 保存 | 印刷 | 拡大

データ | **オプション** | 情報

▼ 検定

裾:

両側検定

*対立仮説: $\mu =$ 60

正規性の仮説

☒ 正規性の検定

ノンパラメトリック検定

☐ 符号検定と Wilcoxon 符号付き順位検定

▼ プロット

プロット:

デフォルトプロット

6. 1標本t検定の結果が表示される。

N	平均	標準偏差	標準誤差	最小値	最大値
10	61.8000	3.0111	0.9522	57.0000	67.0000

平均	平均の 95% 信頼限界		標準偏差	標準偏差の 95% 信頼限界	
61.8000	59.6460	63.9540	3.0111	2.0711	5.4971

自由度	t 値	Pr > t
9	1.89	0.0913

t 値 = 1.89

有意確率 = 0.0913

サンプルサイズが拡大した場合

(B) ランダムに選んだ製品20個の有効成分含有量
母平均：60m g 違いはみられるか？

65	62	57	61	60	58	63	67	62	63
65	62	57	61	60	58	63	67	62	63

20個の平均値を求めると

$$(65+62+\cdots+62+63) \div 20 = 61.8$$

標本平均 (61.8) と母平均 (60.0) の比較

「ワークシート名（Data11）」を入力し、実行ボタンをクリックする。

SAS Studio

サーバーファイルとフォルダ

タスクとユーティリティ

マイタスク

タスク

データ

グラフ

マップ

統計量

データ探索

要約統計量

分布分析

一元度数表

相関分析

分割表分析

検定

線形モデル

生存時間分析

多変量分析

クラスター分析

*プロセスフロー 1 x

プロセスフロー 1 > "EXCELデータ.xlsx" のインポート

設定 コード/結果 分割 実行

オプション ノード

ファイル情報

ソース ファイル

ファイル名: EXCELデータ.xlsx

ソースの場所: /home/u61364448/データ分析の基礎1

ワークシート名:

Data11

出力データ

SAS Server: SASApp

データセット名: IMPORT

ライブラリ: WORK

変更

オプション

ファイルの種類:

デフォルト (ファイル拡張子に基づく)

N	平均	標準偏差	標準誤差	最小値	最大値
20	61.8000	2.9308	0.6553	57.0000	67.0000

平均	平均の 95% 信頼限界		標準偏差	標準偏差の 95% 信頼限界	
61.8000	60.4284	63.1716	2.9308	2.2288	4.2806

自由度	t 値	Pr > t
18	2.75	0.0128

t 値 = 2.75 > 1.89

P値（有意確率） = 0.0128 < 0.0913

効果量 (Cohen's d)

t値 (P値) はサンプルサイズの影響を受ける

$$\begin{aligned} t \text{ 値} &= \frac{\text{差}}{\text{バラツキ}} \times \sqrt{n} \\ &= \text{効果量} \times \sqrt{n} \end{aligned}$$

1標本 t 検定 効果量の比較

	P値	t 値	n	効果量
(A)	0.0913	1.89	10	0.60
(B)	0.0128	2.75	20	0.61

◇効果量（Cohen's d）の大きさの評価

0.2	0.5	0.8
小	中	大

2標本 t 検定

- (A) 新製品の好感度について、男女別各10人に10点満点にて調査した。
男女間の評価に違いは見られるか？

											平均
男性	7	5	8	5	7	3	5	7	6	8	6.1
女性	6	4	7	5	6	3	6	6	4	6	5.3

男性平均（6.1）と女性平均（5.3）の比較

◇ 1 標本 t 検定

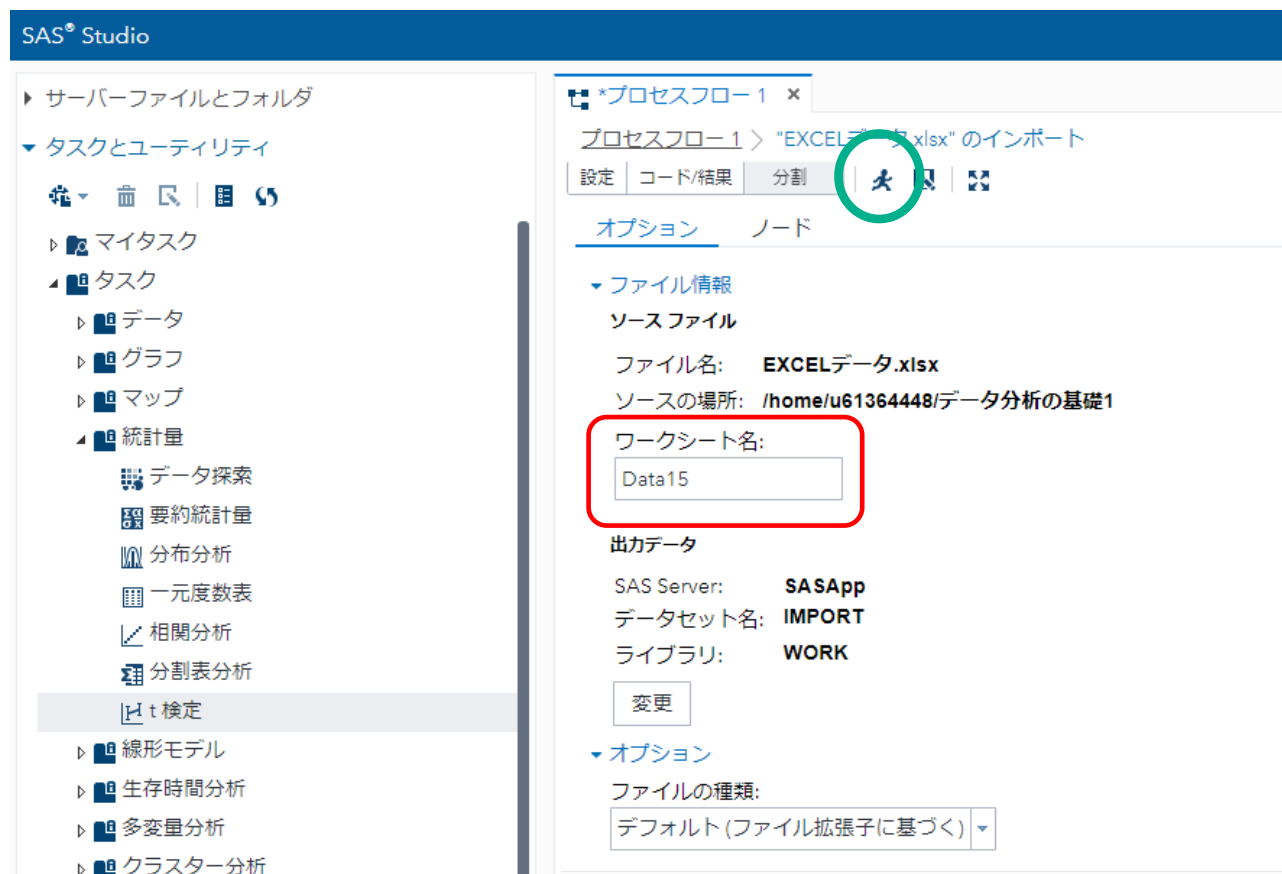
$$t \text{ 値} = \frac{\text{差}}{\text{バラツキ}} \times \sqrt{n}$$

◇ 2 標本 t 検定

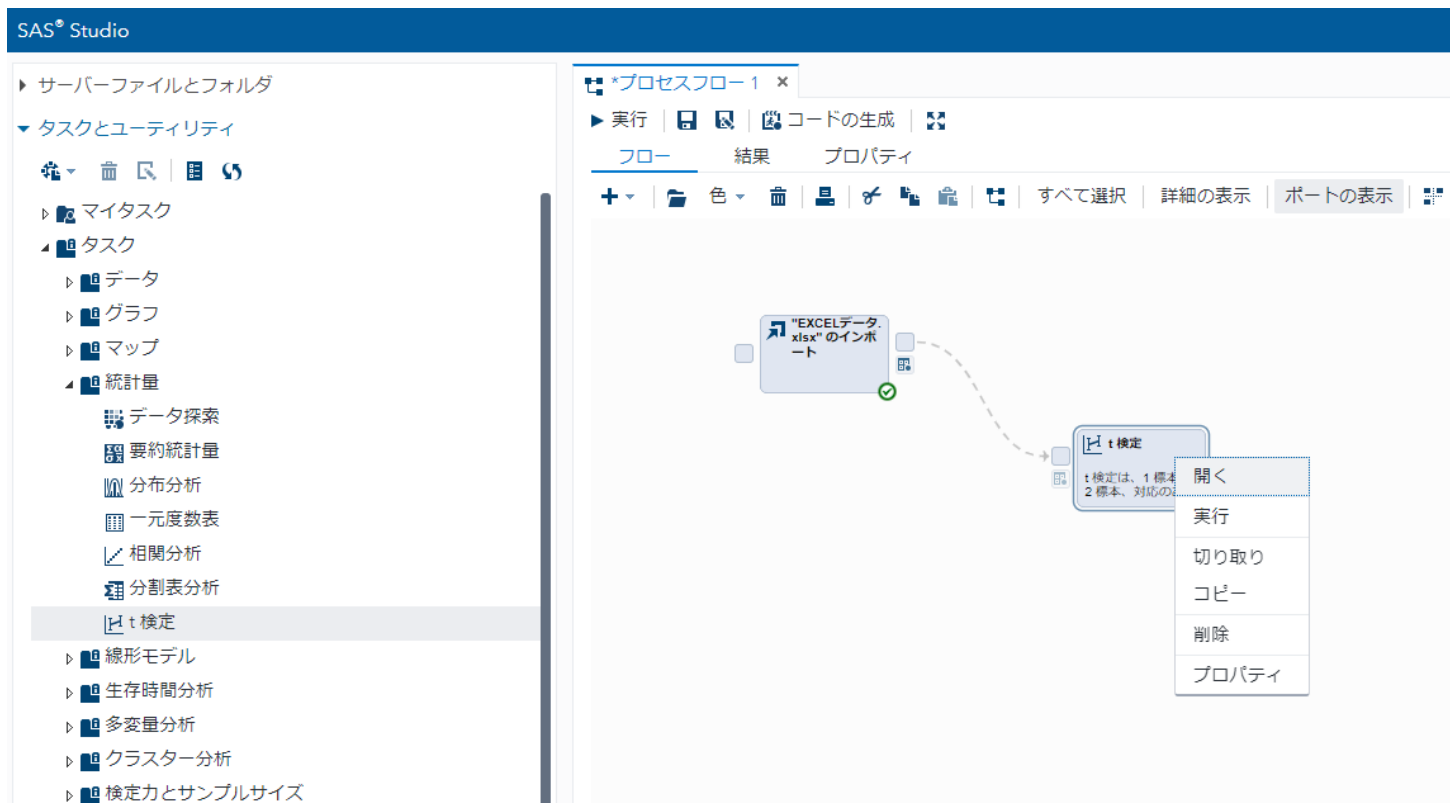
$$t \text{ 値} = \frac{\text{差（男女の平均値の差）}}{\text{男性のバラツキ} + \text{女性のバラツキ}} \times \sqrt{2n}$$

2標本 t 検定 (SAS Studio)

2. 「ワークシート名 (Data15)」を入力し、実行ボタンをクリックする。



3. 「タスクとユーティリティ」→「タスク」→「統計量」の「t 検定」をフォロー画面にドラッグし、「EXCELデータ・・・」と結合、右クリック「開く」をクリックする。



4. 「t検定」を右クリック、開き、「データ」、「t 検定（2標本検定）」、「分析変数（評価）」、「グループ変数（性別）」をセットする。

The screenshot displays the SAS Studio interface. On the left, the 'タスクとユーティリティ' (Tasks and Utilities) pane is expanded, showing the '統計量' (Statistics) category with 't 検定' (t-test) selected. The main workspace shows the 'プロセスフロー 1' (Process Flow 1) task, specifically the 't 検定' (t-test) node. The 'データ' (Data) tab is active, showing 'WORK.IMPORT' as the data source and 'フィルタ: (なし)' (Filter: (none)). The '役割' (Role) section is highlighted with a red rounded rectangle, showing the following configuration:

- t 検定: 2 標本検定
- *分析変数: (1 項目) 評価
- *グループ変数: (1 項目) 性別

5. 結果が表示される。

手法	分散	自由度	t 値	Pr > t
Pooled	Equal	18	-1.25	0.2281
Satterthwaite	Unequal	17.036	-1.25	0.2290

等分散性				
手法	分子の自由度	分母の自由度	F 値	Pr > F
Folded F	9	9	1.62	0.4813

◇分散が違うとは言えないとき
P値（有意確率） = 0.2281

◇分散が違うとき
P値（有意確率） = 0.2290

サンプルサイズが拡大した場合

(B) 新製品の好感度について、男女別各30人に10点満点にて調査した。
男女間の評価に違いは見られるか？

平均											
男性	7	5	8	5	7	3	5	7	6	8	
	7	5	8	5	7	3	5	7	6	8	5.3
	7	5	8	5	7	3	5	7	6	8	
女性	6	4	7	5	6	3	6	6	4	6	6.1
	6	4	7	5	6	3	6	6	4	6	
	6	4	7	5	6	3	6	6	4	6	

男女各10人から男女各30人に増加

「ワークシート名（Data16）」を入力し、実行ボタンをクリックする。

SAS® Studio

サーバーファイルとフォルダ

▼ タスクとユーティリティ

- マイタスク
- タスク
 - データ
 - グラフ
 - マップ
 - 統計量
 - データ探索
 - 要約統計量
 - 分布分析
 - 一元度数表
 - 相関分析
 - 分割表分析
 - t検定**
 - 線形モデル
 - 生存時間分析
 - 多変量分析
 - クラスタ分析

*プロセスフロー 1 ×

プロセスフロー 1 > "EXCELデータ.xlsx" のインポート

設定 | コード/結果 | 分割 | **実行** | 出力

オプション | ノード

▼ ファイル情報

ソース ファイル

ファイル名: EXCELデータ.xlsx

ソースの場所: /home/u61364448/データ分析の基礎1

ワークシート名:
Data16

出力データ

SAS Server: SASApp

データセット名: IMPORT

ライブラリ: WORK

変更

▼ オプション

ファイルの種類:
デフォルト (ファイル拡張子に基づく) ▼

手法	分散	自由度	t 値	Pr > t
Pooled	Equal	58	-2.24	0.0290
Satterthwaite	Unequal	54.895	-2.24	0.0292

等分散性				
手法	分子の自由度	分母の自由度	F 値	Pr > F
Folded F	29	29	1.62	0.1976

P値（有意確率） = 0.0292 < 0.2290

効果量 (Cohen's d)

◇2標本 t 検定

$$\begin{aligned} t \text{ 値} &= \frac{\text{差 (男女の平均値の差)}}{\text{男性のバラツキ} + \text{女性のバラツキ}} \times \sqrt{2n} \\ &= \frac{\text{差 (男女の平均値の差)}}{\text{男性と女性の平均的なバラツキ}} \times \sqrt{\frac{n}{2}} \\ &= \text{効果量} \times \sqrt{\frac{n}{2}} \end{aligned}$$

2標本 t 検定 効果量の比較

	P値	t 値	n	効果量
(A)	0.2290	1.25	10	0.56
(B)	0.0292	2.24	30	0.58

◇効果量（Cohen's d）の大きさの評価

0.2	0.5	0.8
小	中	大

「違い」と「効果」

「違いの大きさ」はサンプルサイズの影響を受ける。

ビッグデータにおいては
「違いがある！」という分析結果が出やすい。

違いがある $\neq$ 効果が大きい

サンプルサイズと効果量を検討することが重要！

まとめ

- 1 データの代表値
平均・バラツキ（分散・標準偏差）
中央値
- 2 データの視覚化
箱ひげ図、ヒストグラム
- 3 正規分布、t 分布
- 4 t 検定
検定のしくみ
P値（有意確率）、有意水準
効果量