

データ分析の基礎-2

2022年10月13日



本日の内容

1.ノンパラメトリック検定
 ウイルコクソン検定

2 サンプルサイズと検出力

3.カイ2乗検定
 フィッシャーの正確検定、効果量
 リスク比とオッズ比

ノンパラメトリック検定

- ・データの値を直接使わず、大きさの順に並べ順位を用いた検定方法

ウイルコクソンの順位和検定

業種と利益額

A	32	11	27	18	33	41	135	42.4
B	42	38	35	34	29	43	51	38.9

データを順位に並べ順位合計を求める。

昇順	11	18	27	29	32	33	34	35	38	41	42	43	51	135
順位	1	2	3	4	5	6	7	8	9	10	11	12	13	14
業種	A	A	A	B	A	A	B	B	B	A	B	B	B	A

Aの順位合計 = $1+2+3+5+6+10+14 = 41$

Bの順位合計 = $4+7+8+9+11+12+13 = 64$

ウイルコクソンの順位和検定の考え方

例) A群6個、B群6個

・順位和のすべての組み合わせについて何通りあるかについて検討する。

A群	1	2	3	4	5	6										順位和
B群							7	8	9	10	11	12				21

順位和が21となるのは1通り

A群	1	2	3	4	5		7									22
B群						6		8	9	10	11	12				

順位和が22となるのは1通り

順位和
23

A群	1	2	3	4	5		8					
B群						6	7	9	10	11	12	

A群	1	2	3	4		6	7					
B群					5		8	9	10	11	12	

23

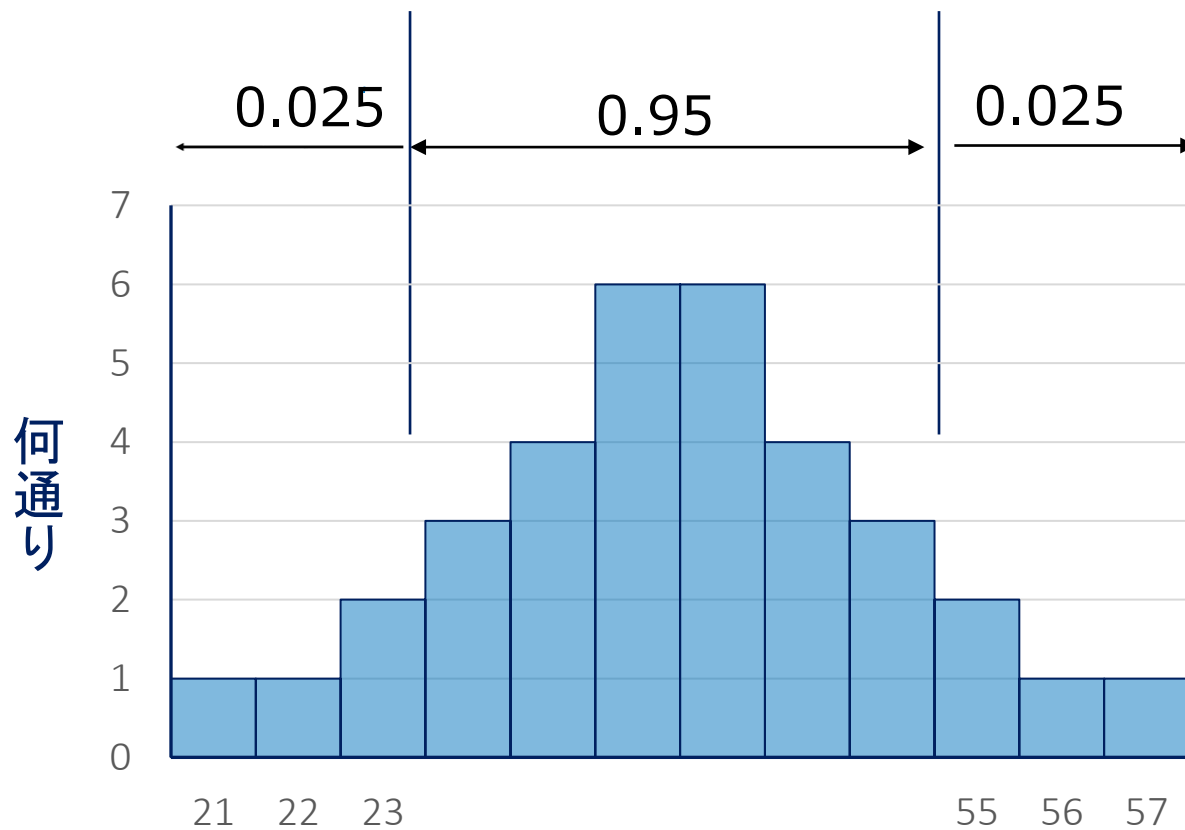
順位和が23となるのは2通り

⋮

A群	7	8	9	10	11	12						
B群							1	2	3	4	5	6

57

順位和が57となるのは1通り



全通りを100として該当する順位和のグラフ位置から判定する。

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 the SAS OnDemand for Academics [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 'Actions: [Clear my saved tabs.](#)'. The background is a dark blue gradient with a pattern of small white dots.

Notices

SAS® OnDemand for Academics
Sign In

SAS Profile email address or user ID

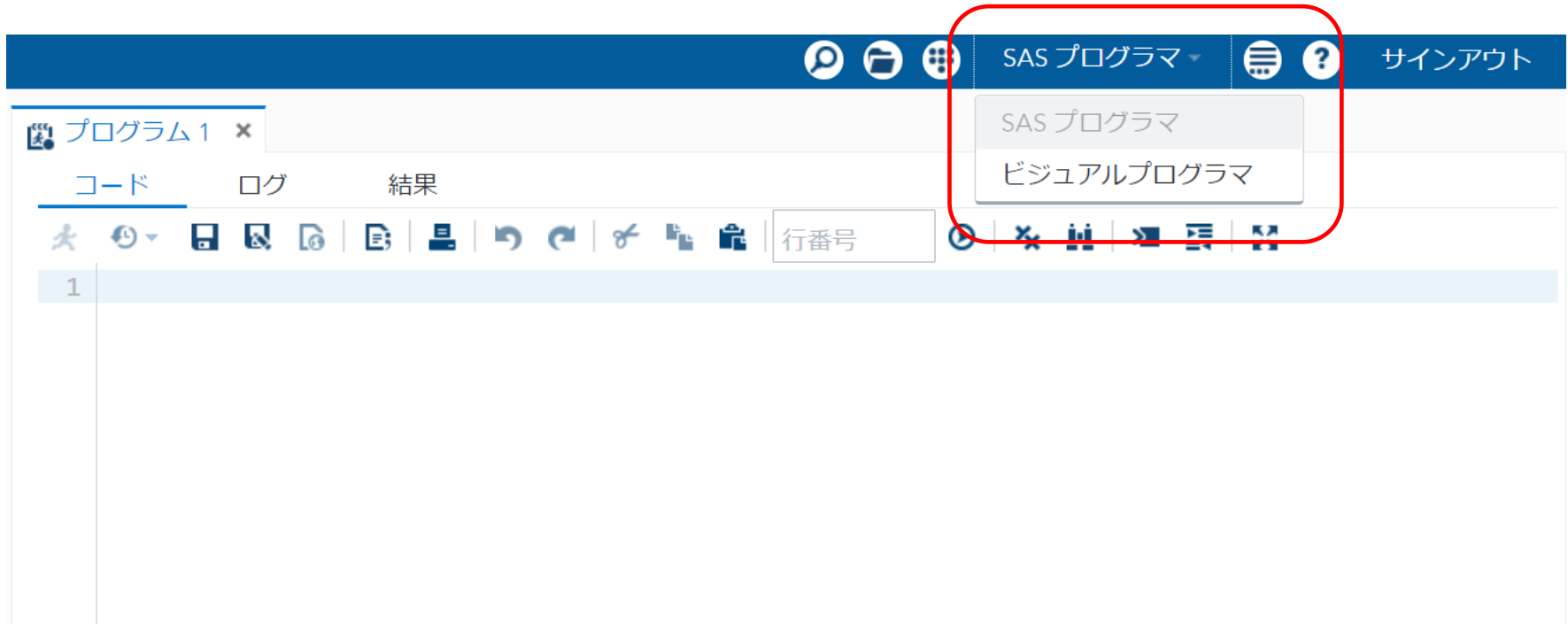
Password

☐ Accept the terms of the SAS OnDemand for Academics [use and conditions](#)

Applications Enrollments Courses

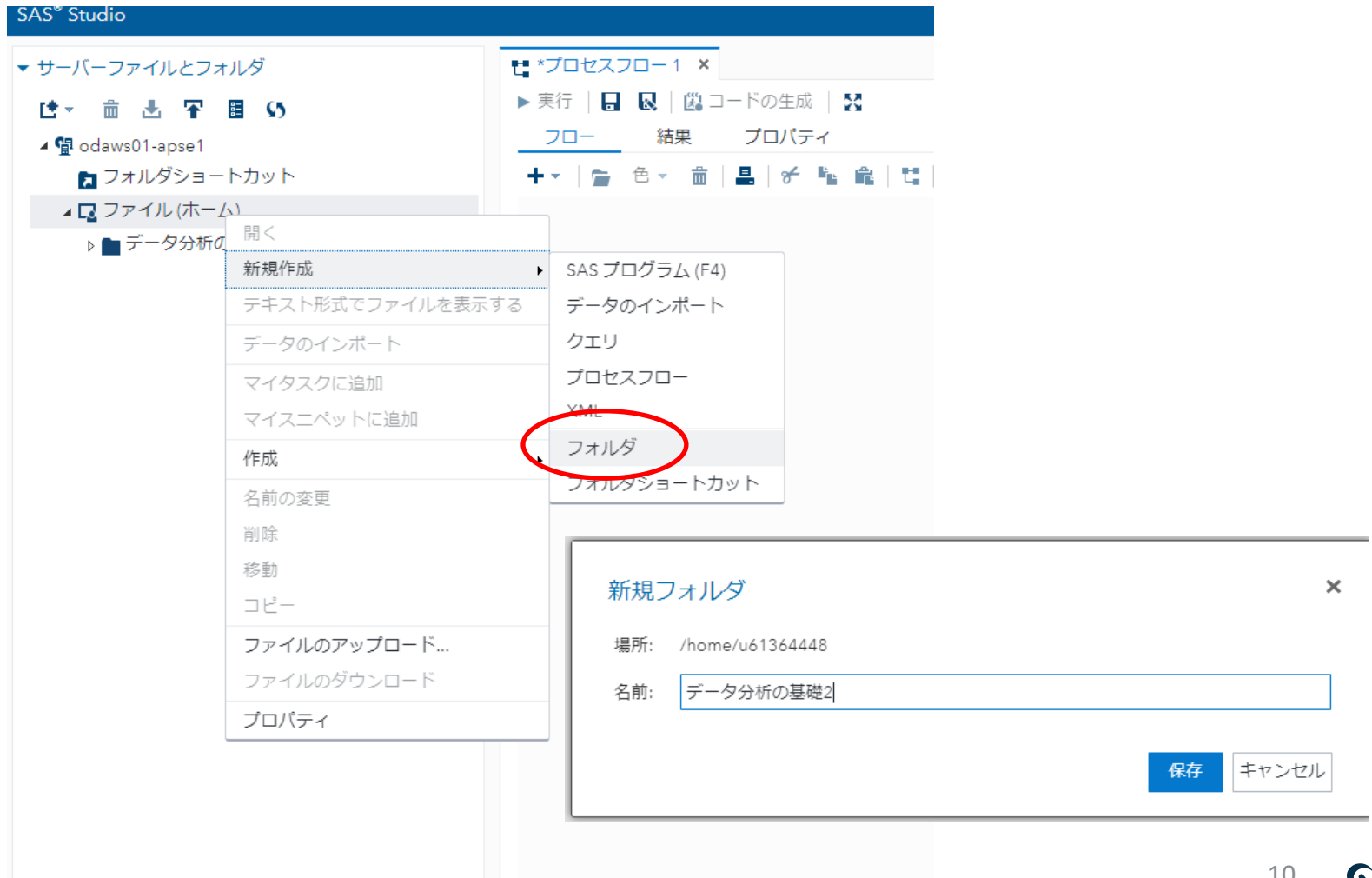
 **SAS® Studio**
Write and run SAS code with a Web-based SAS development environment.
Actions: [Clear my saved tabs.](#)

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。



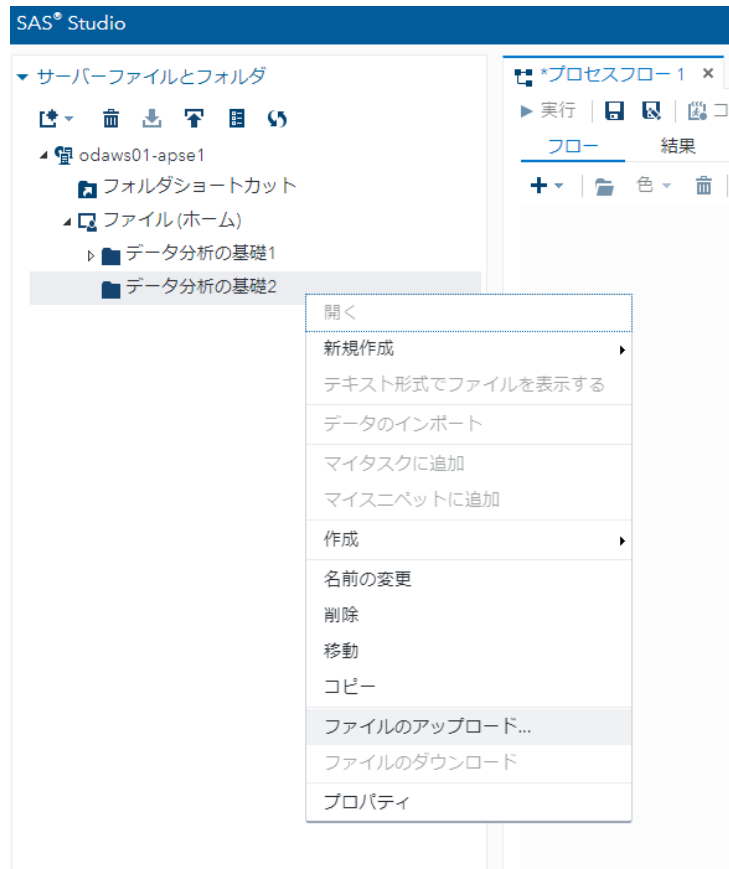
4. 新規フォルダを作成する。

「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2）」を入力し、「保存」をクリックする。

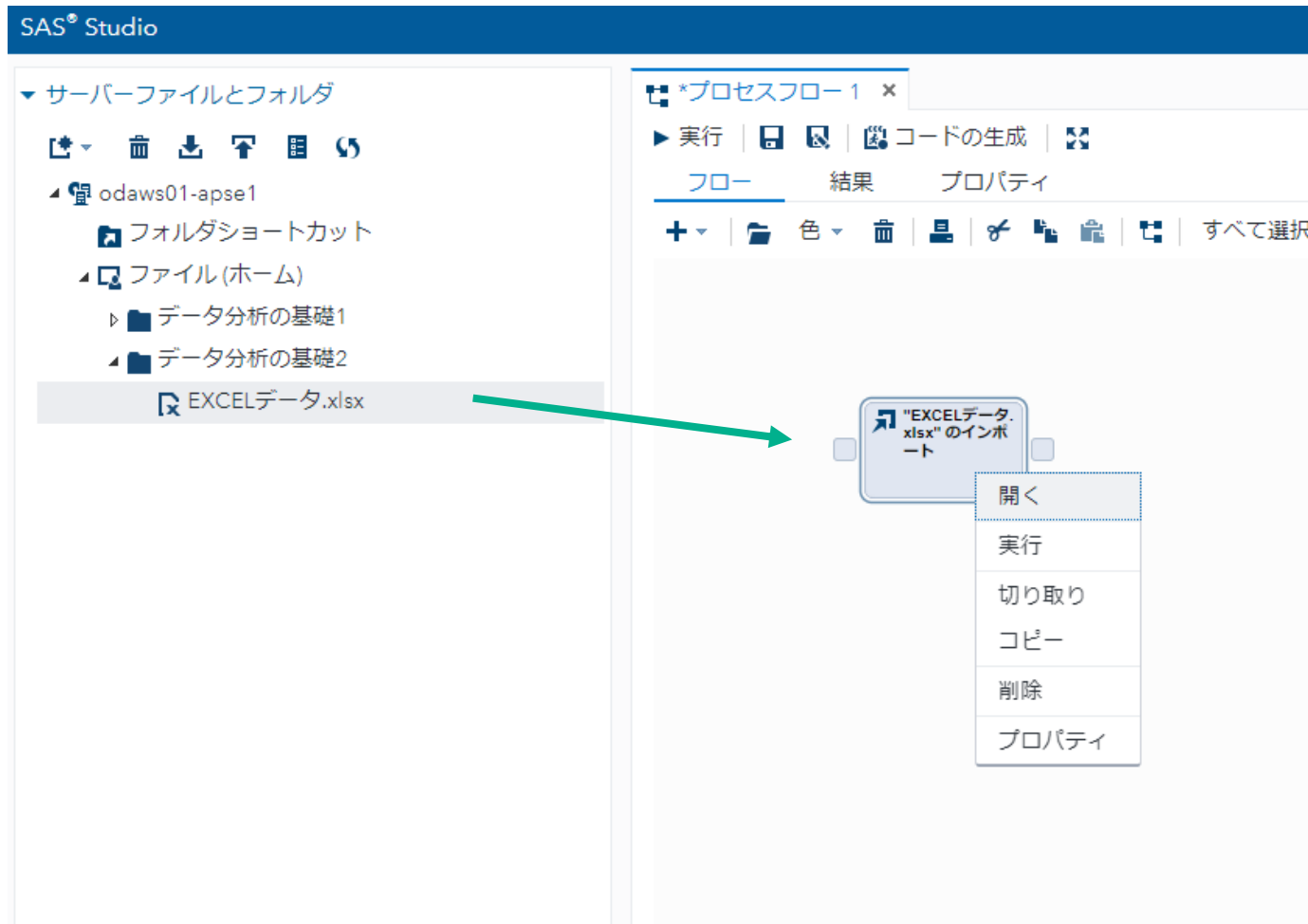


5. ファイルをアップロードする。

「データ分析の基礎2」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2」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
 - データ分析の基礎2
 - 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** | 詳細

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2

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

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。

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
 - データ分析の基礎2
 - 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 | 🔍 | 🔄

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2

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

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)

列

- ☒ すべて選択
- ☒ ▲ 業種
- ☒ 📄 利益額

プロパティ | 値

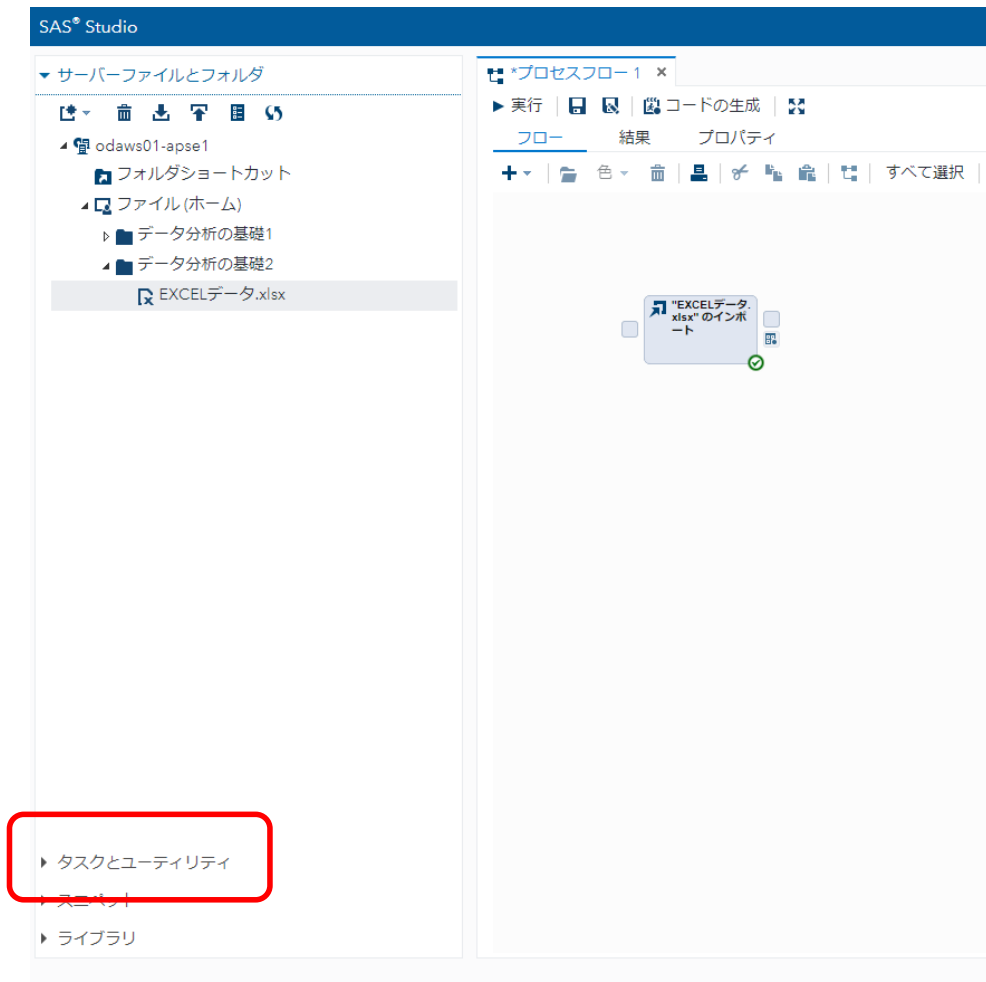
合計行数: 14 合計列数: 2

	業種	
1	A	
2	A	
3	A	
4	A	
5	A	

タスクとユーティリティ

スニペット

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. 「タスクとユーティリティ」→「タスク」→「統計量」の「t 検定」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 'タスク' (Tasks) > '統計量' (Statistics) > 't 検定' (t-test). The main workspace shows a flowchart with a task named '"EXCELデータ.xlsx" のインポート' (Import "EXCEL data.xlsx"). A dashed arrow points from this task to a 't 検定' (t-test) task. A right-click context menu is open over the 't 検定' 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).

11. 「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（利益額）」、「グループ変数（業種）」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 shows the '統計量' (Statistics) category expanded, with 't 検定' (t-test) selected. The main workspace shows a process flow with a node named 't 検定'. The 'データ' (Data) tab is active, showing 'WORK.IMPORT1' as the data source and 'フィルタ: (なし)' (Filter: (none)). The '役割' (Role) tab is also visible, showing the configuration for the 't 検定' node. The 't 検定' dropdown is set to '2 標本検定' (2-sample t-test). The '分析変数: (1 項目)' (Analysis variable: (1 item)) is set to '利益額' (Profit). The 'グループ変数: (1 項目)' (Group variable: (1 item)) is set to '業種' (Industry). The '役割' tab and its contents are highlighted with a red rounded rectangle.

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
- クラスター分析
- 検定力とサンプルサイズ
- 統計的工程管理
- 組み合わせと確率
- データマイニング
- 計量経済
- 予測

*プロセスフロー 1 ×

プロセスフロー 1 > t 検定

設定 | コード/結果 | 分割

データ | オプション | 情報 | ノード

データ

ワーク: WORK.IMPORT1

フィルタ: (なし)

役割

t 検定:

2 標本検定

*分析変数: (1 項目)

123 利益額

*グループ変数: (1 項目)

業種

12. 「オプション」をクリックし、「Wilcoxon順位和検定」をチェックする。

The screenshot shows the SAS Studio interface. On the left, the 'Tasks and Utilities' pane is expanded to 'Statistics', where 't-test' is selected. The main window displays the 'Options' tab for the 't-test' procedure. Under the 'Test' section, the 'Test:' dropdown is set to 'Two-sided test'. The null hypothesis is set to $\mu_1 - \mu_2 = 0$. The 'Normality test' checkbox is checked. The 'Nonparametric test' section is expanded, and the 'Wilcoxon Rank Sum Test' checkbox is checked and highlighted with a red box. The 'Plots' section shows the 'Plot:' dropdown set to 'Default plot'.

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
- クラスター分析
- 検定力とサンプルサイズ
- 統計的工程管理
- 組み合わせと確率
- データマイニング

*プロセスフロー 1 x

プロセスフロー 1 > t 検定

設定 | コード/結果 | 分割 | 実行 | 出力 | 印刷

データ | オプション | 情報 | ノード

検定

検定:

両側検定

*対立仮説: $\mu_1 - \mu_2 =$ 0

☐ 不等分散に対する Cox and Cochran 確率近似

正規性の仮説

☒ 正規性の検定

ノンパラメトリック検定

☒ Wilcoxon 順位和検定

プロット

プロット:

デフォルトプロット

13. Wilcoxon順位和検定の結果が表示される。

変数 利益額 に対する Wilcoxon スコア (順位和) 分類変数: 業種					
業種	N	スコアの 合計	H0 のもとでの 期待値	H0 のもとでの 標準偏差	平均 スコア
A	7	41.0	52.50	7.826238	5.857143
B	7	64.0	52.50	7.826238	9.142857

Wilcoxon の順位和検定 (2 標本)					
統計量	Z	Pr < Z	Pr > Z	t 近似	
				Pr < Z	Pr > Z
41.0000	-1.4055	0.0799	0.1599	0.0917	0.1833
Z には 0.5 の連続性の補正が含まれています。					

P値 (有意確率) = 0.1599

2標本t検定の場合

手法	分散	自由度	t 値	Pr > t
Pooled	Equal	12	0.22	0.8283
Satterthwaite	Unequal	6.3519	0.22	0.8315

$$P\text{値} = 0.8315 > 0.1599$$

ノンパラメトリック検定のまとめ

・データの値を直接使わず、大きさの順に並べ順位を用いた検定方法

- ・外れ値がある場合は有効性が高い。
- ・母集団の分布の形に関わらず有効である。
- ・母集団が正規分布を仮定できる場合、パラメトリック検定の方が検出力が高くなることが多い。

サンプルサイズと検出力（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

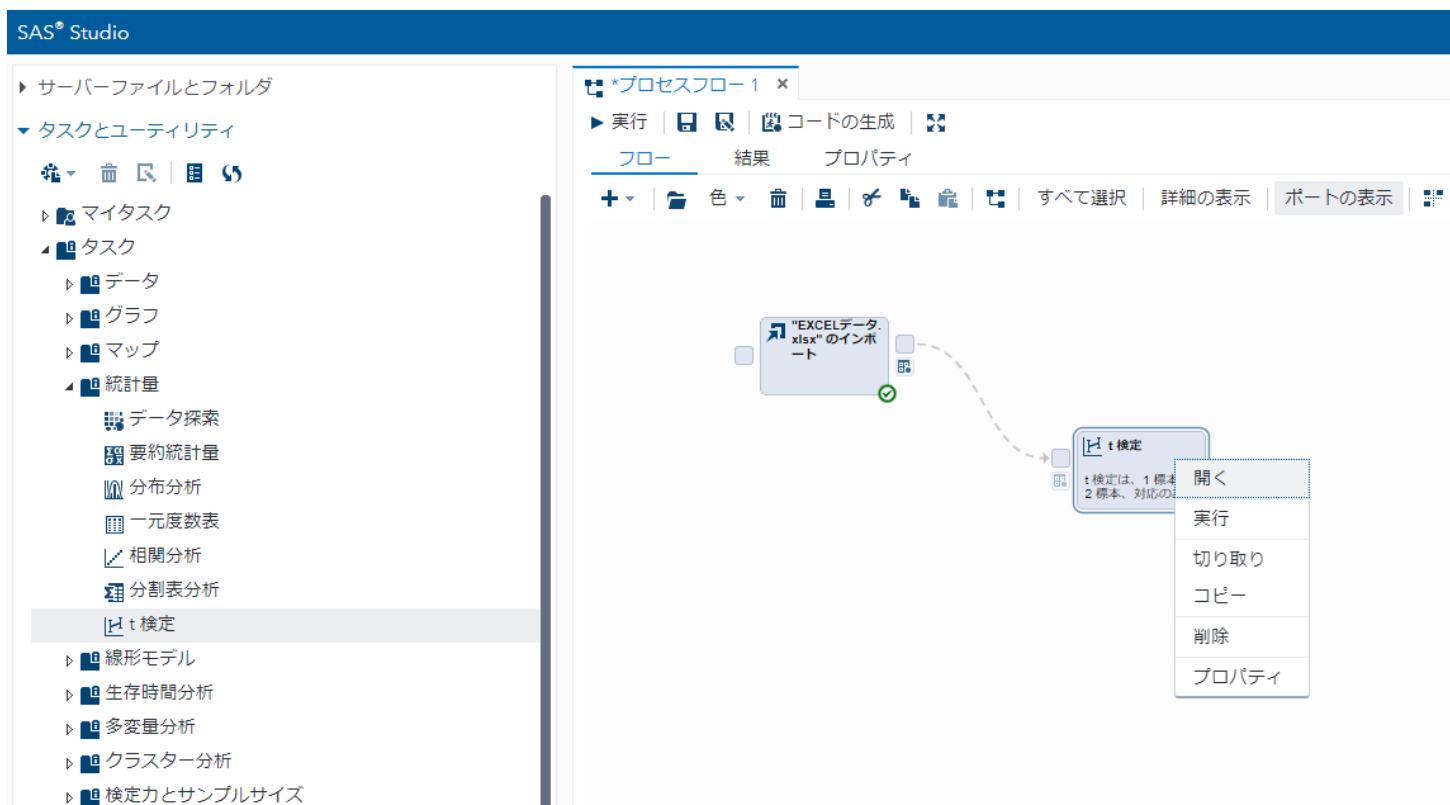
男性平均（6.1）と女性平均（5.3）の比較

2標本 t 検定 (SAS Studio)

1. 「ワークシート名 (Data3)」を入力し、実行ボタンをクリックする。



2. 「タスクとユーティリティ」→「タスク」→「統計量」の「t 検定」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。



3. 「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 'タスク' (Tasks) section with '統計量' (Statistics) selected. Under '統計量', 't 検定' (t-test) is highlighted. On the right, the 'プロセスフロー' (Process Flow) pane shows the 't 検定' (t-test) node selected. 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 expanded, showing 't 検定: 2 標本検定' (t-test: 2-sample test). Below this, the '分析変数' (Analysis Variable) is set to '評価' (Evaluation) and the 'グループ変数' (Grouping Variable) is set to '性別' (Gender). The '役割' section is highlighted with a red rounded rectangle.

結果が表示される。

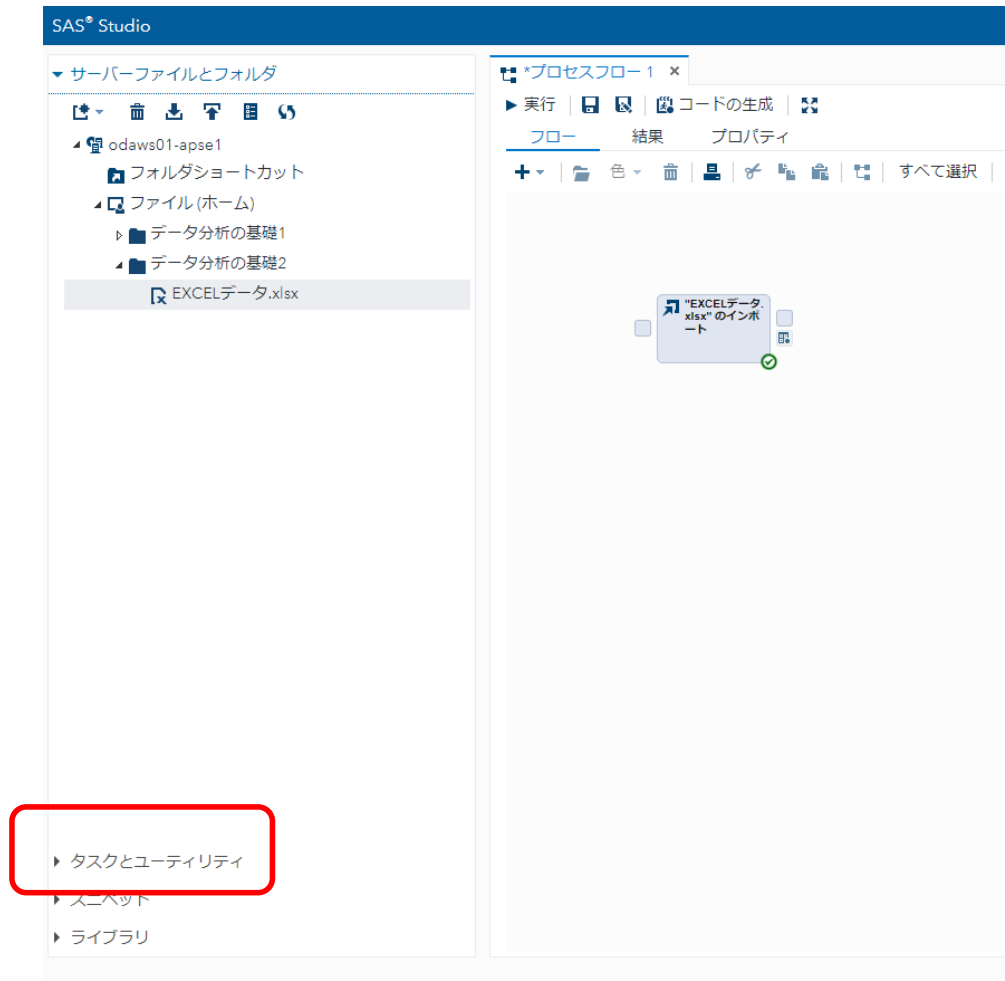
性別	手法	平均	平均の 95% 信頼限界		標準偏差	標準偏差の 95% 信頼限界	
f		5.3000	4.4046	6.1954	1.2517	0.8609	2.2851
m		6.1000	4.9589	7.2411	1.5951	1.0972	2.9121
Diff (1-2)	Pooled	-0.8000	-2.1471	0.5471	1.4337	1.0833	2.1202
Diff (1-2)	Satterthwaite	-0.8000	-2.1526	0.5526			

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

効果量 : 0.560

検出力を求める

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



2. 「タスクとユーティリティ」- 「検定力とサンプルサイズ」- 「t 検定」をフロー画面にドラッグし、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 open, showing a list of statistical tasks. The '検定力とサンプルサイズ' (Power and Sample Size) category is expanded, and the 't 検定' (t-test) task is highlighted with a red rectangle. On the right, the 'フロー' (Flow) pane shows a process flow. A task named 'EXCELデータ.xlsx'のインポート (Import Excel data.xlsx) is connected to a 't 検定' task. A context menu is open over the 't 検定' task, with the '開く' (Open) option selected. The menu also includes options for '実行' (Run), '切り取り' (Cut), 'コピー' (Copy), '削除' (Delete), and 'プロパティ' (Properties).

3. 各項目を設定し、検定力（検出力）を求める。

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

- 変換統計量
- 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
- クラスター分析
- 検定力とサンプルサイズ
 - Pearson 相関
 - 多重回帰
 - 信頼区間
 - 比率の検定
 - t 検定
 - 一元配置分散分析
 - ロジスティック回帰
 - 生存ランク検定
 - Cox 回帰
 - Wilcoxon 検定
 - 同等性検定
 - カスタム検定
- 統計的工程管理
- 組み合わせと確率
- データマイニング
- 計量経済

*プロセスフロー 1 x

プロセスフロー 1 > t 検定 1

設定 | コード/結果 | 分割 | 実行 | 保存 | 印刷

プロパティ | プロット | 情報 | ノード

検定の種類

t 検定の種類: 2 標本検定

解決

☒ 検定力

☐ 合計サンプルサイズ

☐ 各グループのサンプルサイズ

分析の詳細

データの仮定分布の選択:

☐ 対数正規分布

☒ 正規分布

検定の境界の数の選択: 両側検定

検定の選択:

☐ プールされた t 検定

☒ Satterthwaite t 検定

有意水準

アルファ値: (最低 1 行) 0.05

平均

フォームの選択: 平均の差

平均差の値: (最低 1 行) 0.8

- ▶ ログヘッディング出力
- ▶ 生存ランク検定
- ▶ Cox 回帰
- ▶ Wilcoxon 検定
- ▶ 同等性検定
- ▶ カスタム検定
- ▶ 統計的工程管理
- ▶ 組み合わせと確率
- ▶ データマイニング
- ▶ 計量経済
- ▶ スニペット
- ▶ ライブラリ

▼ 標準偏差

グループの標準偏差の値: (最低 1 行)

[タイトルなし]



グループ 1

グループ 2

1.5951

1.2517

▼ サンプルサイズ

☐ 小数値のサンプルサイズを許可する

フォームの選択:

各グループのサンプルサイズ



各グループのサンプルサイズ: (最低 1 行)



10

検出力が求められる。

不等分散の平均差に対する 2 標本 t 検定

固定シナリオ要素	
分布	Normal
手法	Exact
裾の数	2
名目の有意水準	0.05
平均差	0.8
群 1 の標準偏差	1.595
群 2 の標準偏差	1.25166
群ごとの標本サイズ	10
帰無仮説の差	0

計算された 検出力	
実際の有意水準	検出力
0.0488	0.215

検出力 : 0.215

◇検定における判断の誤り

①「男女の好感度が同じとき」

男女の好感度が「違う」と判定してしまう誤り

(例) P値 (危険率) = 0.03

男女の好感度が「違う」と判定してしまう誤りは3%
= 100回「違う」と判定して3回間違える。



あわてものの誤り (第1種の過誤)

習慣として 0.05 (5%) がよく使用される

②「男女の好感度が違うとき」
「違うとは言えない」と判定してしまう誤り
(見過ごす誤り)



ぼんやりものの誤り (第2種の過誤)

一般に0.20がよく使われる。
100回判定して20回間違える。(見過ごす。)
= 100回判定して80回検出する。

⇒ 検出力 = 0.8

検出力の基準として 0.80が比較的よく使用される

サンプルサイズを求める。

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
- クラスター分析
- 検定力とサンプルサイズ
 - Pearson 相関
 - 多重回帰
 - 信頼区間
 - 比率の検定
 - t 検定
 - 一元配置分散分析
 - ロジスティック回帰

プロセスフロー 1 x t 検定 x

設定 | コード/結果 | 分割 | 実行 | 保存 | 印刷 | 共有

プロパティ | プロット | 情報

検定の種類

t 検定の種類: 2 標本検定

解決

☐ 検定力

☒ 合計サンプルサイズ

☐ 各グループのサンプルサイズ

分析の詳細

データの仮定分布の選択:

☐ 対数正規分布

☒ 正規分布

検定の限界の数の選択: 両側検定

検定の選択:

☐ プールされた t 検定

☒ Satterthwaite t 検定

有意水準

アルファ値: (最低 1 行) 0.05

平均

フォームの選択:

平均の差

平均差の値: (最低 1 行) 0.8

- ▷ 多変量分析
- ▷ クラスター分析
- ▲ 検定力とサンプルサイズ
 - ≡ Pearson 相関
 - ※ 多重回帰
 - ≡ 信頼区間
 - ≡ 比率の検定
 - ≡ t 検定
 - ≡ 一元配置分散分析
 - ≡ ロジスティック回帰
 - ≡ 生存ランク検定
 - ≡ Cox 回帰
- ▷ スニペット

▼ 標準偏差

グループの標準偏差の値: (最低 1 行)



グループ 1

グループ 2

1.5951

1.2517

▼ 検定力

検定力値: (最低 1 行)

0.8

▼ サンプルサイズ

☐ 小数値のサンプルサイズを許可する

相対グループサイズの選択:

☒ グループサイズは等しい

☐ グループサイズは等しくない

サンプルサイズが求められる。

不等分散の平均差に対する 2 標本 t 検定

固定シナリオ要素	
分布	Normal
手法	Exact
裾の数	2
名目の有意水準	0.05
平均差	0.8
群 1 の標準偏差	1.5951
群 2 の標準偏差	1.2517
名目の検出力	0.8
帰無仮説の差	0
群 1 の重み	1
群 2 の重み	1

サンプルサイズ : 104

計算された 総標本サイズ		
実際の有意水準	実際の検出力	総標本サイズ
0.05	0.804	104

検定結果の検討

P値、効果量、検出力について検討する。

1) P値（有意確率、危険率）

標本サイズに比例して t値は大きくなる。
= 標本サイズに比例してP値は小さくなる。

P値が小さい $\neq$ 効果が大い

2) 効果量（標本サイズの影響を受けない）

3) 検出力（標本サイズは充分かどうか）

標本サイズが小さいとき：検出力が充分か検討する。
標本サイズが大きいとき：効果量を重視する。

効果量・検出力・サンプルサイズ・P値（危険率）の関係



- ・効果量が大きいとき、サンプルサイズは小
- ・効果量が小さいとき、サンプルサイズは大

- ・サンプルサイズが大きいとき、検出力は大
- ・サンプルサイズが小さいとき、検出力は小

分割表の検定（カイ二乗検定）

主力商品 A、B 2種類について、一般層と富裕層に
A、B どちらを選択するかについて調査した。違いは見られるか。

(人)

	A	B	計
一般層	60	40	100
富裕層	30	70	100
計	90	110	200

実現値と期待値の食い違いを求める

<期待値>

(人)

	A	B	計
一般層	50	50	100
富裕層	50	50	100
計	90	110	200

(人)

	A	B	計
一般層	45	55	100
富裕層	45	55	100
計	90	110	200

$$\text{食い違い} = \frac{(\text{実現値} - \text{期待値})^2}{\text{期待値}} \text{の合計}$$

$$\begin{aligned} & \frac{(60 - 45)^2}{45} + \frac{(40 - 55)^2}{55} \\ & \frac{(30 - 45)^2}{45} + \frac{(70 - 55)^2}{55} \\ & = 18.18 \end{aligned}$$

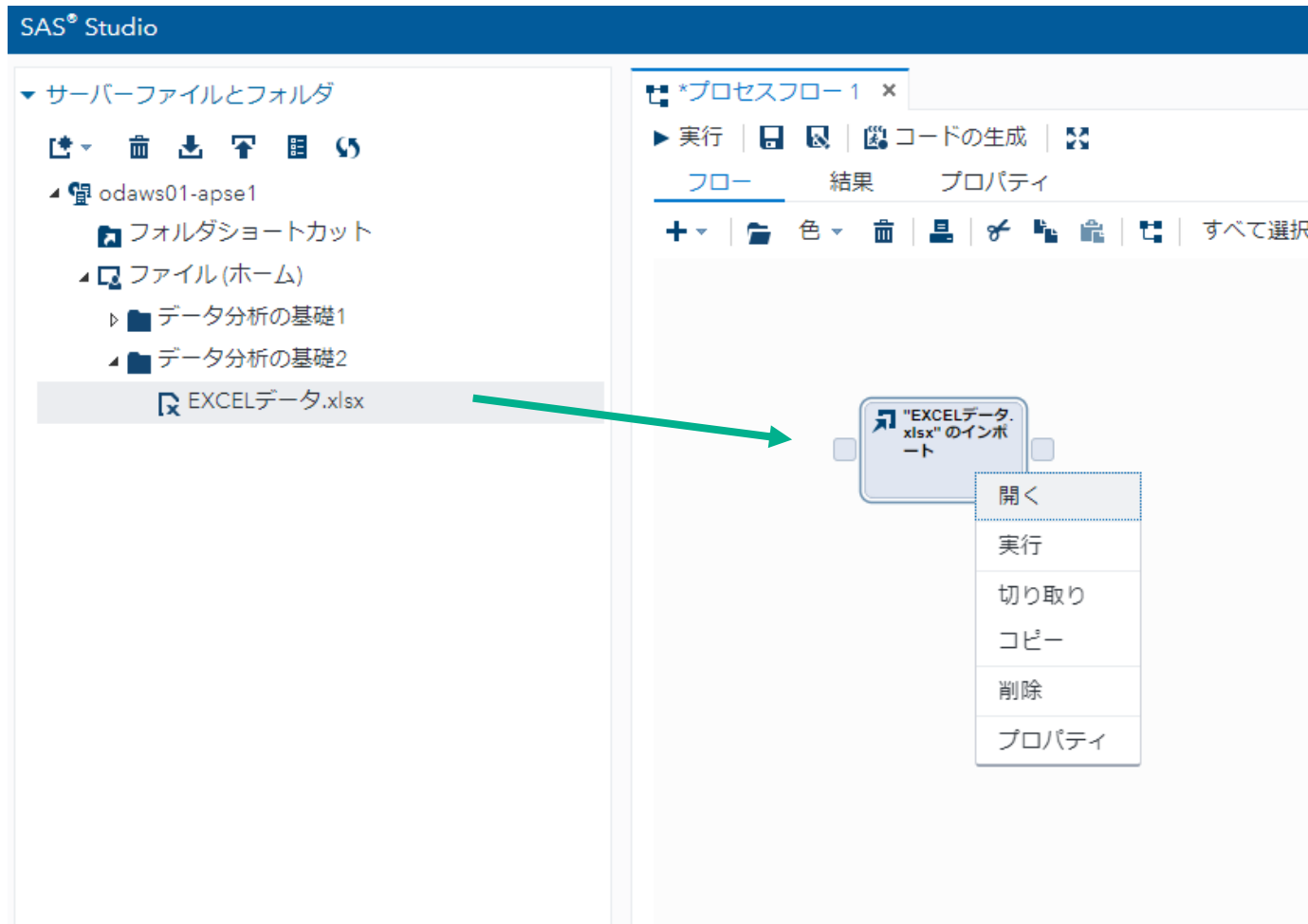
カイ2乗分布

分割表の検定

データを作成する。

	A	B	C	D
1	層	商品	N	
2	1	1	60	
3	1	2	40	
4	2	1	30	
5	2	2	70	
6				
7				

1. 「データ分析の基礎2」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。

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
 - クラスター分析
 - 検定力とサンプルサイズ

*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** | 中止 | 再試行

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2

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

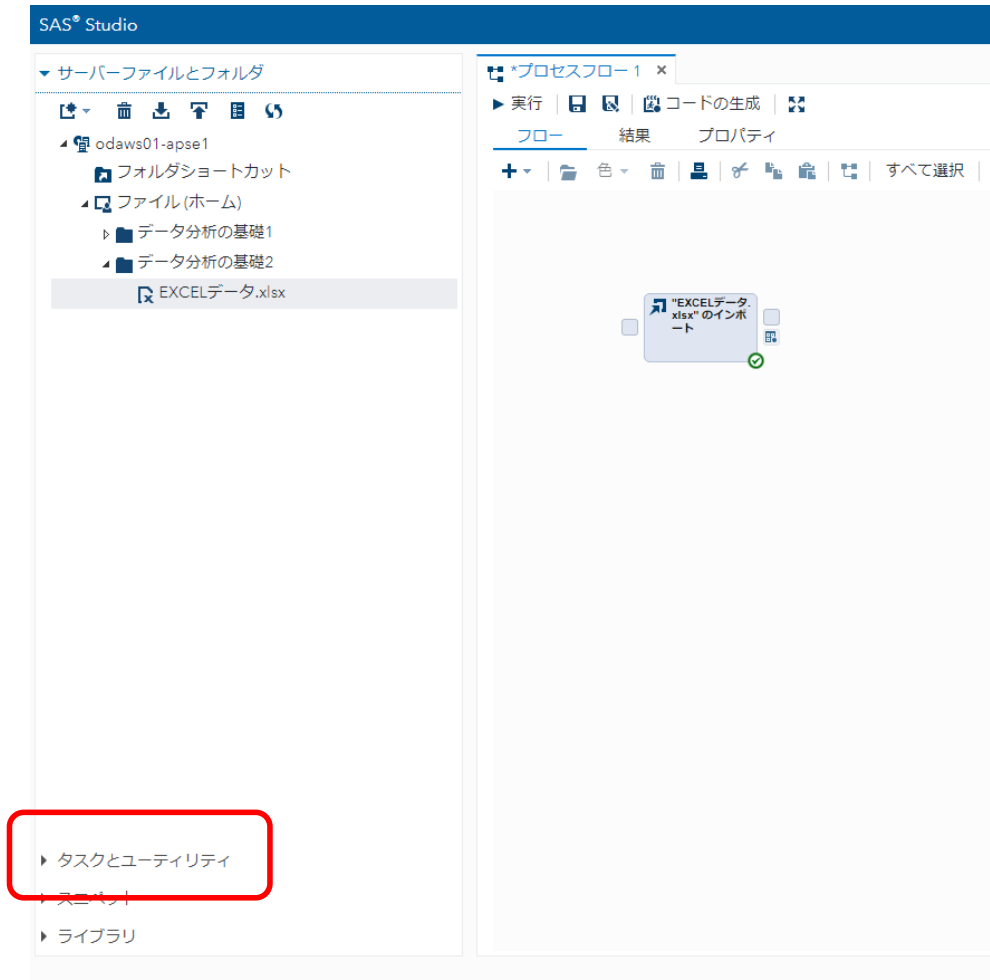
オプション

ファイルの種類:

デフォルト (ファイル拡張子に基づく) ▼

☒ SAS 変数名の生成

3. 「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。



4. 「タスクとユーティリティ」→「タスク」→「統計量」の「分割表分析」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, showing the '統計量' (Statistics) category. The '分割表分析' (Contingency Analysis) task is highlighted. The main workspace shows a process flow with two tasks: 'EXCELデータ.xlsxのインポート' (Import Excel Data.xlsx) and '分割表分析' (Contingency Analysis). 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 '分割表分析' task, with the '開く' (Open) option selected.

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
 - クラスター分析
 - 検定力とサンプルサイズ

*プロセスフロー 1 x

実行 | コードの生成

フロー | 結果 | プロパティ

+ | 色 | すべて選択 | 詳細の表示 | ポートの

"EXCELデータ.xlsx"のインポート

分割表分析

開く

実行

切り取り

コピー

削除

プロパティ

5. 「役割」－「行変数」に「層」、「列変数」に「商品」を設定する。

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
 - クラスター分析
 - 検定力とサンプルサイズ
 - 統計的工程管理
 - 組み合わせと確率
 - データマイニング
 - 計量経済
 - 予測
 - ネットワーク最適化
- ユーティリティ
 - データのインポート
 - プロセスフロー

*プロセスフロー 1

プロセスフロー 1 > 分割表分析

設定 | コード/結果 | 分割 | 実行 | 保存 | 印刷

データ | オプション | 情報 | ノード

データ

WORK.IMPORT

フィルタ: (なし)

役割

行変数:

- 層

列変数:

- 商品

層の変数:

- 列

6. 「追加役割」をクリックし、「度数カウント」に「N」を設定し、実行する。

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
 - クラスター分析
 - 検定力とサンプルサイズ
 - 統計的工程管理
 - 組み合わせと確率
 - データマイニング
 - 計量経済
 - 予測
 - ネットワーク最適化
- ユーティリティ
 - データのインポート
 - プロセスフロー
- スニペット
- ニゾニ

*プロセスフロー 1 ×

プロセスフロー 1 > 分割表分析

設定 | コード/結果 | 分割 | **実行** | 情報

データ | オプション | 情報 | ノード

列変数:

- 商品

層の変数:

- 列

追加役割

度数カウント: (1 項目)

- N

統計量	自由度	値	p 値
カイ 2 乗値	1	18.1818	<.0001
尤度比カイ 2 乗値	1	18.4803	<.0001
連続性補正カイ 2 乗値	1	16.9899	<.0001
Mantel-Haenszel のカイ 2 乗値	1	18.0909	<.0001
ファイ係数		0.3015	
一致係数		0.2887	
Cramer の V 統計量		0.3015	

カイ二乗値 : 18.1818
P値 : < 0.0001

サンプルサイズが小さい場合

主力商品 A、B 2種類について、一般層と富裕層に
A、B どちらを選択するかについて調査した。違いは見られるか。

(人)

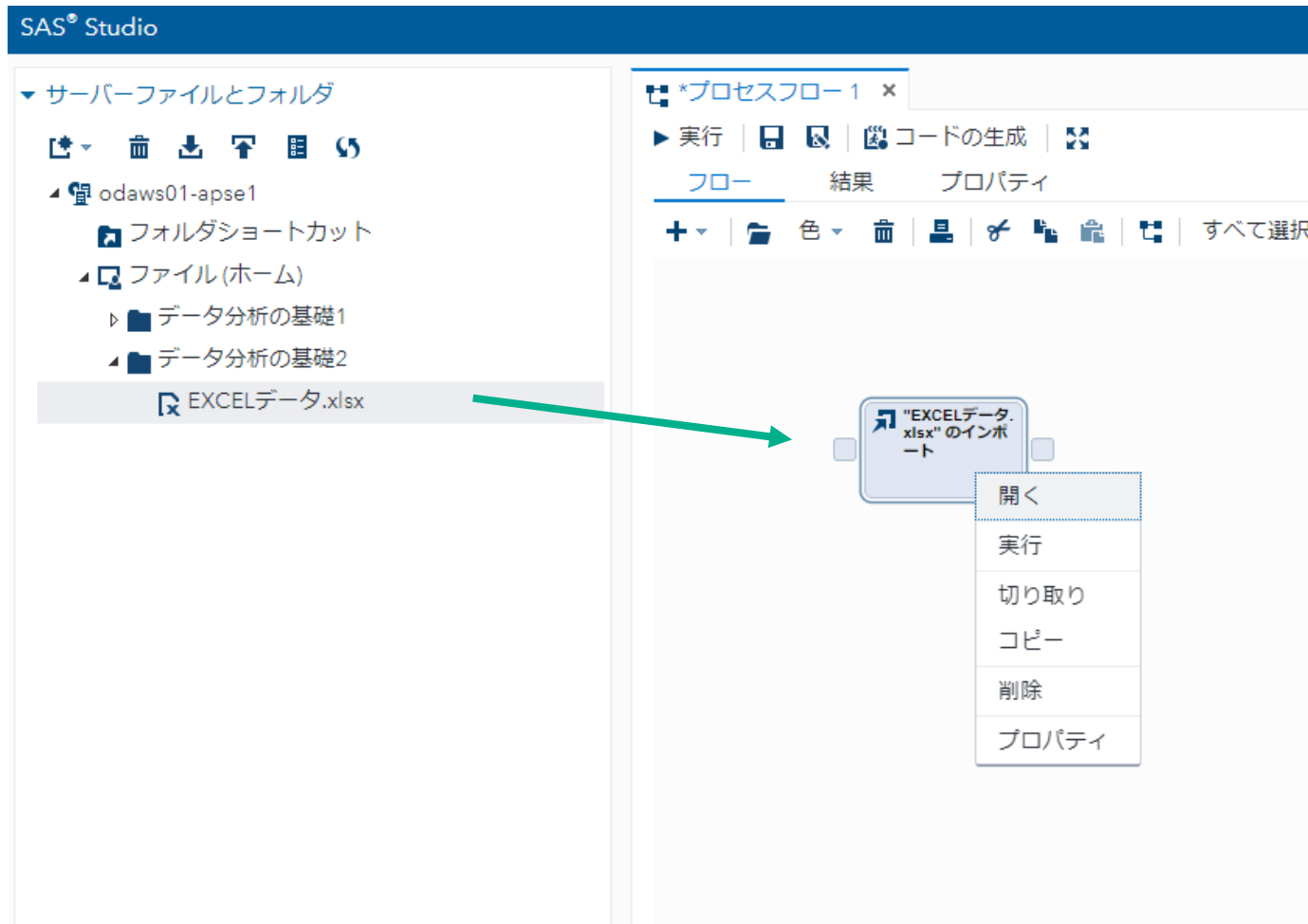
	A	B	計
一般層	6	4	10
富裕層	3	7	10
計	9	11	20

分割表の検定

データを作成する。

	A	B	C	D	E
1	層	商品	N		
2	1	1	6		
3	1	2	4		
4	2	1	3		
5	2	2	7		
6					
7					

1. 「データ分析の基礎2」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（Data6）」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)
 - sasuser.v94
 - データ分析の基礎1
 - データ分析の基礎2

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** | 中止 | 再試行

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2

ワークシート名:
Data6

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) ▼

☒ SAS 変数名の生成

統計量	自由度	値	p 値
カイ 2 乗値	1	1.8182	0.1775
尤度比カイ 2 乗値	1	1.8480	0.1740
連続性補正カイ 2 乗値	1	0.8081	0.3687
Mantel-Haenszel のカイ 2 乗値	1	1.7273	0.1888
ファイ係数		0.3015	
一致係数		0.2887	
Gramer の V 統計量		0.3015	

カイ二乗値 : 1.8182
P値 : 0.1775

期待値が5未満のセルが20%以上ある、または1未満が1つでもある場合に使用

⇒ フィッシャーの正確検定

SAS Studio フィッシャーの正確検定

「オプション」－「正確検定」－「Fisherの正確検定」をチェックする。

The screenshot displays the SAS Studio interface. On the left is a navigation pane with a tree structure. The 'Statistics' (統計量) folder is expanded, and 'Contingency Analysis' (分割表分析) is selected. Below it, 'Fisher's Exact Test' (Fisherの検定) is highlighted. On the right is the 'Options' (オプション) tab for the selected analysis. This tab contains several expandable sections: 'Degrees of Freedom' (自由度), 'Percent' (パーセント), 'Cumulative' (累積), 'Chi-Square' (カイ2乗), 'Statistics' (統計量), and 'Exact Test' (正確検定). The 'Exact Test' section is expanded, and the checkbox for 'Fisher's Exact Test' (Fisherの正確検定) is checked. A red rectangle highlights this checkbox. At the bottom of the 'Exact Test' section, there is a note in Japanese: '注: 大規模な問題の一部では、正確検定の計算に大量の時間とメモリが必要になる場合があります。' (Note: In some cases of large-scale problems, a large amount of time and memory may be required for the calculation of the exact test.)

Left Panel (Navigation):

- 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**
 - Fisherの検定
 - 線形モデル
 - 生存時間分析
 - 多変量分析
 - クラスター分析
 - 検定力とサンプルサイズ
 - 統計的工程管理
 - 組み合わせと確率
 - データマイニング
 - 計量経済
 - 予測
 - ネットワーク最適化
- ユーティリティ
 - データのインポート
 - プログラマフロー
- スニペット

Right Panel (Options):

- データ
- オプション**
- 情報
- ノート
- ▼ 度数表
 - ▼ 度数
 - ☒ 観測
 - ☐ 期待
 - ☐ 偏差
 - ▼ パーセント
 - ☐ セル
 - ☐ 行
 - ☐ 列
 - ▼ 累積
 - ☐ 列のパーセント
 - ☐ 度数とパーセント
 - ▼ カイ2乗
 - ☐ カイ2乗統計量に対するセルの寄与率
 - ▼ 統計量
 - ☒ カイ2乗統計量
 - ☐ 関連性の指標
 - ☐ Cochran-Mantel-Haenszel 統計量
 - ☐ 一致の統計量 (正方形な表)
 - ☐ オッズ比と相対リスク (2x2 表)
 - ☐ 二項比率とリスクの差 (2x2 表)
 - ▼ 正確検定**
 - ☒ Fisherの正確検定

注: 大規模な問題の一部では、正確検定の計算に大量の時間とメモリが必要になる場合があります。

フィッシャーの正確検定

標本サイズ = 200

統計量	自由度	値	p 値
カイ 2 乗値	1	18.1818	<.0001
尤度比カイ 2 乗値	1	18.4803	<.0001
連続性補正カイ 2 乗値	1	16.9899	<.0001
Mantel-Haenszel のカイ 2 乗値	1	18.0909	<.0001
ファイ係数		0.3015	
一致係数		0.2887	
Cramer の V 統計量		0.3015	

Fisher の正確検定	
セル (1,1) 度数 (F)	60
左側 Pr <= F	1.0000
右側 Pr >= F	<.0001
表の確率 (P)	<.0001
両側 Pr <= P	<.0001

標本サイズ = 200

標本サイズ = 20

統計量	自由度	値	p 値
カイ 2 乗値	1	1.8182	0.1775
尤度比カイ 2 乗値	1	1.8480	0.1740
連続性補正カイ 2 乗値	1	0.8081	0.3687
Mantel-Haenszel のカイ 2 乗値	1	1.7273	0.1888
ファイ係数		0.3015	
一致係数		0.2887	
Cramer の V 統計量		0.3015	
WARNING: セルの50%において、期待度数が5より小さくなっています。 カイ2乗検定は妥当な検定でないと思われます。			

Fisher の正確検定	
セル (1,1) 度数 (F)	6
左側 Pr <= F	0.9651
右側 Pr >= F	0.1849
表の確率 (P)	0.1500
両側 Pr <= P	0.3698

標本サイズ = 20

分割表の検定における効果量：ファイ係数

$$\begin{aligned}\text{ファイ係数} &= \sqrt{\frac{\text{カイ二乗値}}{n}} \\ &= \sqrt{\frac{18.18}{200}} = \sqrt{\frac{1.818}{20}} = 0.302\end{aligned}$$

◇効果量の大きさの評価

0.1	0.3	0.5
小	中	大

リスク比

	不整脈有無		横計	割合
	ある	ない		
喫煙	3	2	5	0.6
非喫煙	1	4	5	0.2

原因は左側。結果は上側とする。

喫煙者が不整脈となるリスク : $3 \div 5 = 0.6$
非喫煙者が不整脈となるリスク : $1 \div 5 = 0.2$

$$\frac{\text{喫煙者のリスク}}{\text{非喫煙者のリスク}} = \frac{0.6}{0.2} = 3$$

「喫煙者が不整脈となるリスクは非喫煙者に比べ3倍」

オッズ比

	不整脈有無		横計	割合
	ある	ない		
喫煙	3	2	5	0.6
非喫煙	1	4	5	0.2
割合	3.0	0.5		

不整脈のある場合の割合 : $3 \div 1 = 3$

不整脈のない場合の割合 : $2 \div 4 = 0.5$

オッズ比 : $3 \div 0.5 = 6$

≠

「喫煙者が不整脈となるリスクは非喫煙者に比べ6倍」

コホート研究とケースコントロール研究

◇コホート研究

- 1) 不整脈がない人に、今までの喫煙の有無を調査。
- 2) その後の2年間、喫煙の有無別に不整脈の発生を追跡調査。

（2年後への研究であり、「前向き」の研究。）

◇ケースコントロール研究

- ・不整脈があると診断された200人と健常者200人に、過去の喫煙の有無を調査。

（過去への研究であり、「後ろ向き」の研究。）

◇ケースコントロール（後ろ向き）研究1

	不整脈有無		横計	割合
	ある	ない		
喫煙	94	74	168	0.56
非喫煙	106	126	232	0.46
割合	0.89	0.59	400	

・不整脈があると診断された200人とランダムに選んだ健康者200人の過去の喫煙の有無を調査。

リスク比	1.22	(0.56/0.46)
オッズ比	1.51	(0.89/0.59)

◇ケースコントロール（後ろ向き）研究2

	不整脈有無		横計	割合
	ある	ない		
喫煙	94	37	131	0.72
非喫煙	106	63	169	0.63
割合	0.89	0.59	300	

・不整脈があると診断された200人とランダムに選んだ健康者100人の過去の喫煙の有無を調査。

リスク比	1.14	(0.72/0.63)
オッズ比	1.51	(0.89/0.59)

◇割合が小さいとき

	不整脈有無		横計	割合
	ある	ない		
喫煙	30	970	1000	0.03
非喫煙	10	990	1000	0.01
割合	3.00	0.98	2000	

リスク比 3.00 (0.03/0.01)

オッズ比 3.06 (3.00/0.98)

割合が小さいとき、リスク比はオッズ比で近似できる。

- ・リスク比は解釈しやすいが、ケースコントロール（後ろ向き）研究では間違った解釈となりやすい。
- ・オッズ比はよく使用される。
割合が小さいとき、リスク比に近似する。

まとめ

- ・ノンパラメトリック検定
 ウィルコクソン順位和検定
- ・検出力とサンプルサイズ
- ・カイ二乗検定
- ・リスク比とオッズ比