

統計学(基礎)

第2回

データの準備と出力の利用

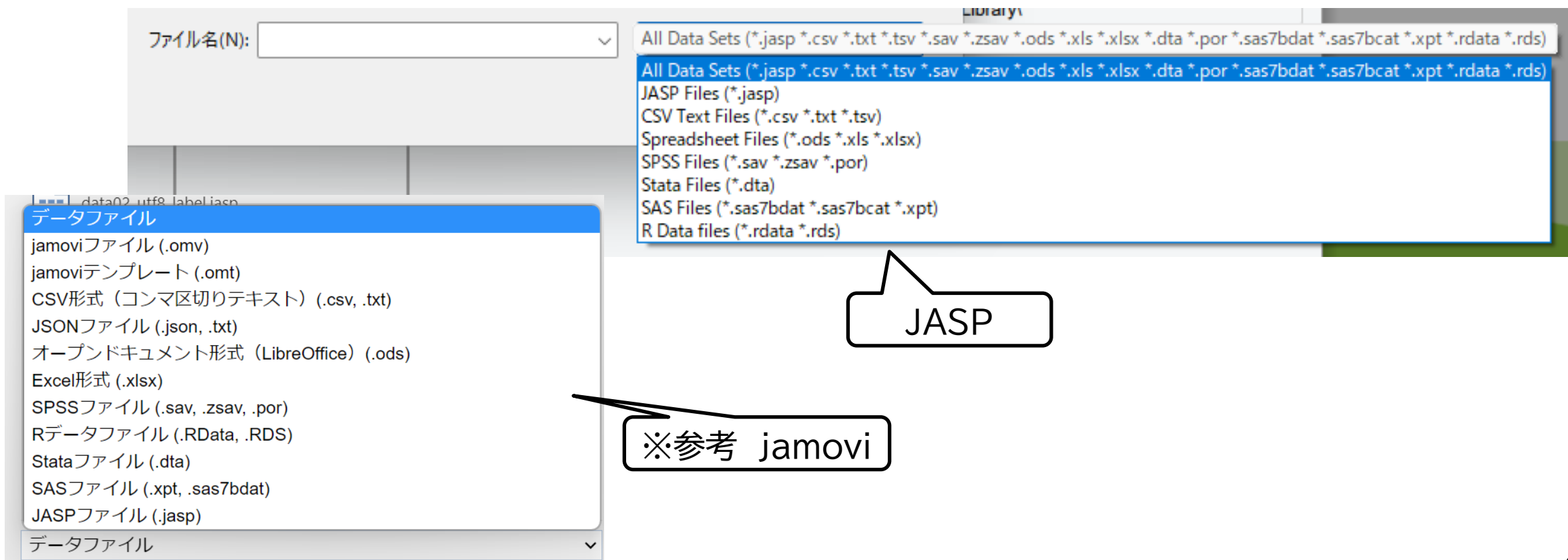
JASPで扱えるデータについて

データライブラリの場所

- データの場所
- C:\Program Files (x86)\JASP\resources\Data Sets\Data Library
- データ単体はCSV形式
- JASPで保存するとJASP形式になる
 - 出力や設定とセットのファイルになる

JASPで読み込めるデータ形式

- 一応いろいろ読める



実際に読んでみると

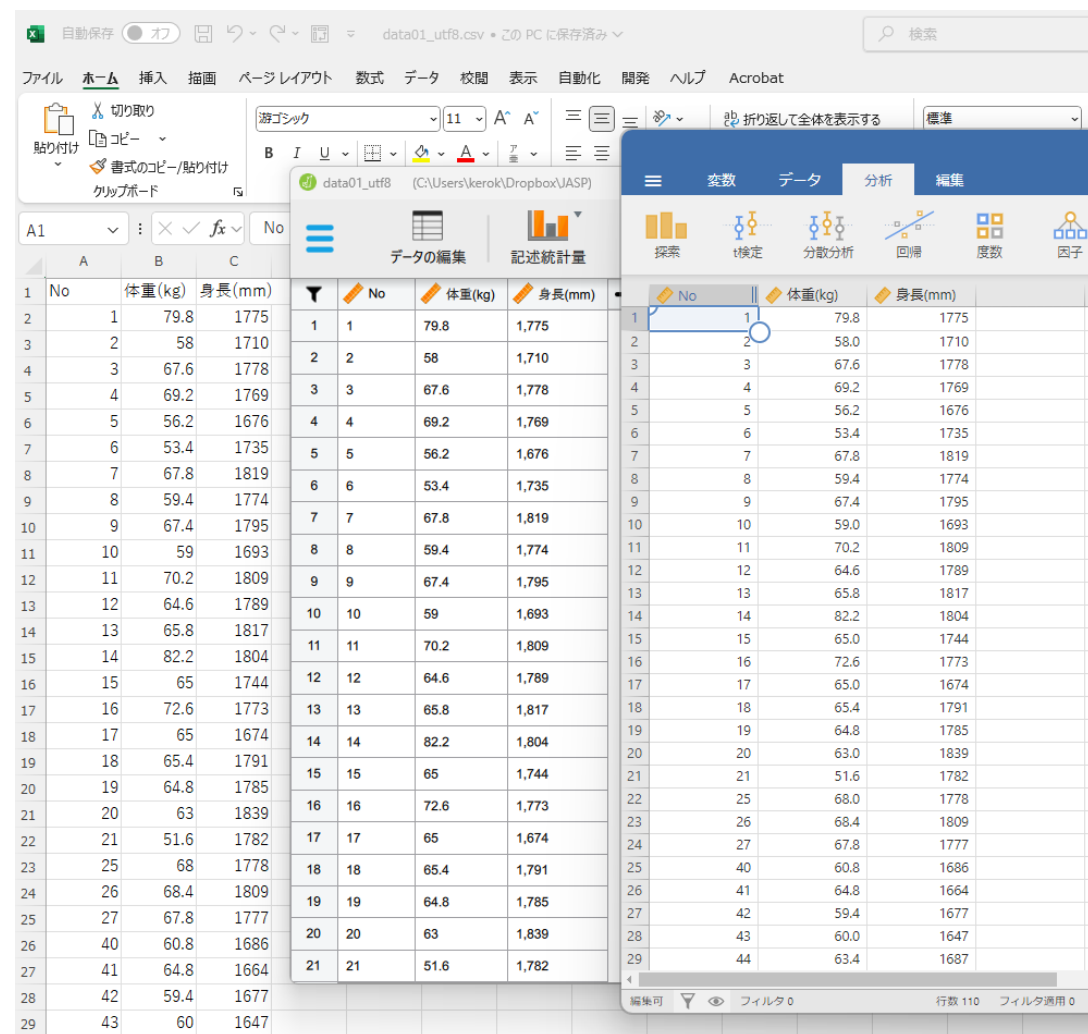
- CSV
– (Shift-JIS)
- 左から
Excel
JASP
jamovi
- タイトルの日本語が読めない

The image shows three overlapping windows displaying the same CSV data. The data is as follows:

No	体重(kg)	身長(mm)
1	79.8	1775
2	58	1710
3	67.6	1778
4	69.2	1769
5	56.2	1676
6	53.4	1735
7	67.8	1819
8	59.4	1774
9	67.4	1795
10	59	1693
11	70.2	1809
12	64.6	1789
13	65.8	1817
14	82.2	1804
15	65	1744
16	72.6	1773
17	65	1674
18	65.4	1791
19	64.8	1785
20	63	1839
21	51.6	1782
22	68	1778
23	68.4	1809
24	67.8	1777
25	60.8	1686
26	64.8	1664
27	60.0	1647
28	59.4	1677

実際に読んでみると

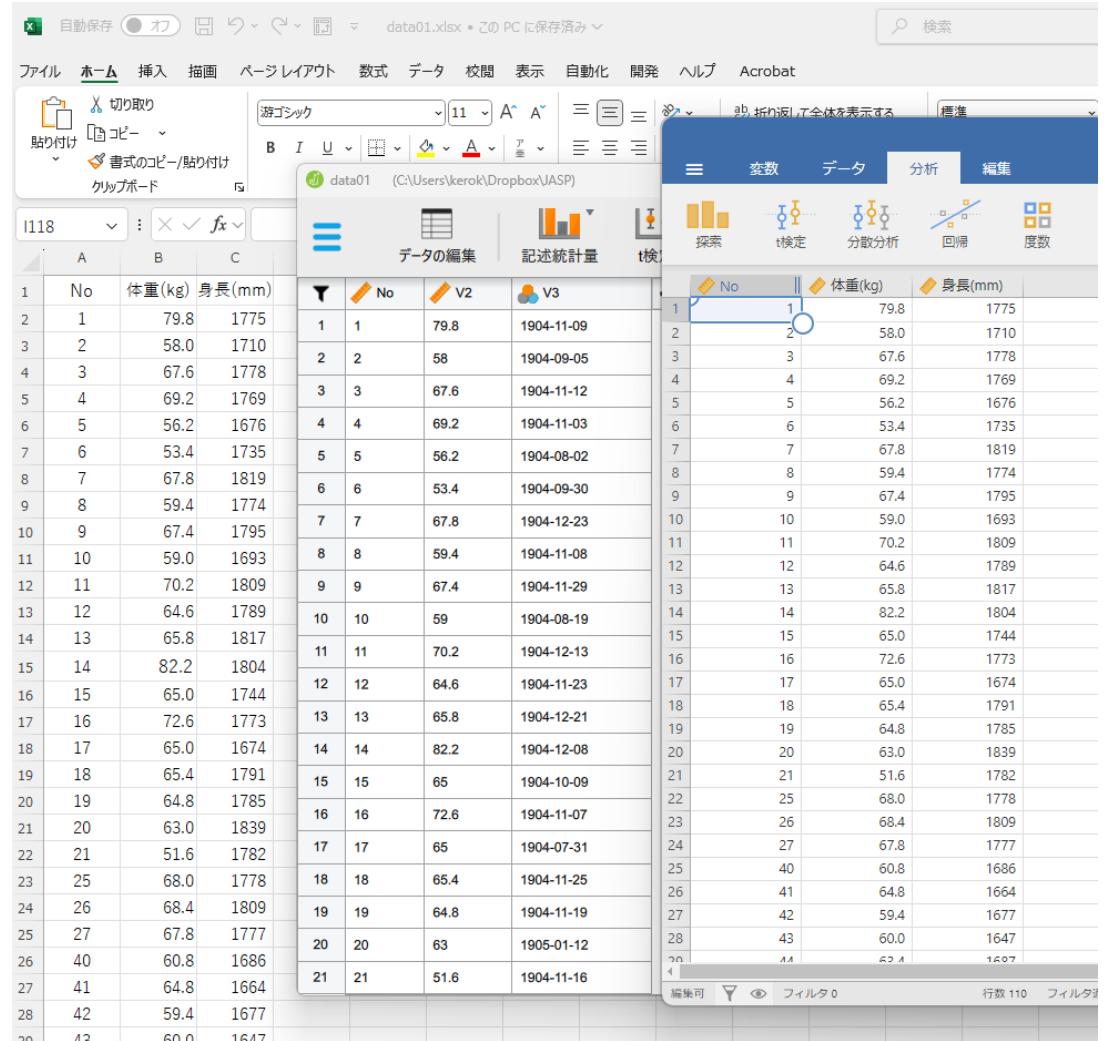
- CSV UTF-8
- ほぼ問題なし



No	体重(kg)	身長(mm)
1	79.8	1775
2	58	1710
3	67.6	1778
4	69.2	1769
5	56.2	1676
6	53.4	1735
7	67.8	1819
8	59.4	1774
9	67.4	1795
10	59	1693
11	70.2	1809
12	64.6	1789
13	65.8	1817
14	82.2	1804
15	65	1744
16	72.6	1773
17	65	1674
18	65.4	1791
19	64.8	1785
20	63	1839
21	51.6	1782
22	68	1778
23	68.4	1809
24	67.8	1777
25	60.8	1686
26	64.8	1664
27	59.4	1677
28	60	1647
29	60	1647

実際に読んでみると

- Excel
- JASPはきついのかも
(Macだと問題ないことも多い)

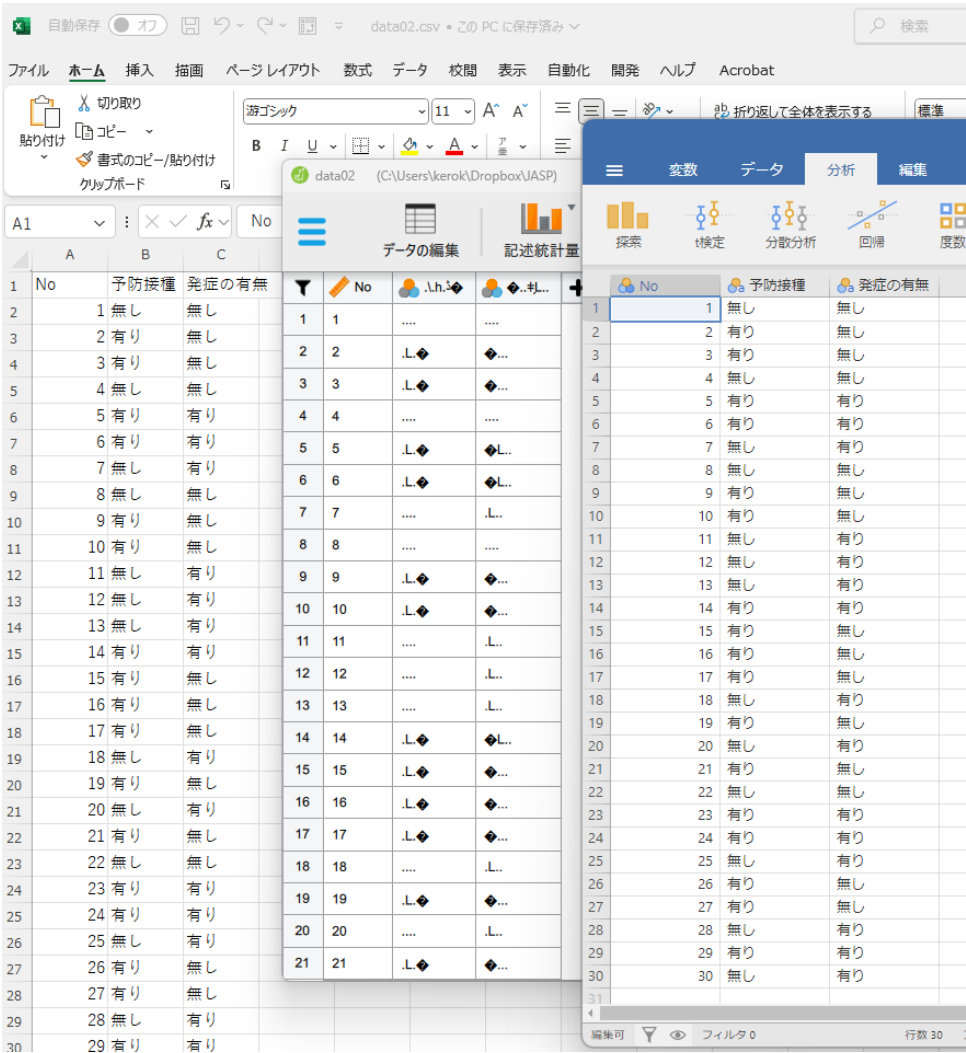


The screenshot displays the JASP software interface. The main window shows a data table with the following columns: No, 体重(kg), and 身長(mm). The data is organized into a grid with rows numbered 1 to 28. The interface includes a menu bar at the top with options like 'ファイル', 'ホーム', '挿入', '描画', 'ページレイアウト', '数式', 'データ', '校閲', '表示', '自動化', '開発', 'ヘルプ', and 'Acrobat'. A sidebar on the right contains various analysis options such as '変数', 'データ', '分析', and '編集'. The bottom status bar indicates '編集可', 'フィルタ 0', '行数 110', and 'フィルタ'.

No	体重(kg)	身長(mm)
1	79.8	1775
2	58.0	1710
3	67.6	1778
4	69.2	1769
5	56.2	1676
6	53.4	1735
7	67.8	1819
8	59.4	1774
9	67.4	1795
10	59.0	1693
11	70.2	1809
12	64.6	1789
13	65.8	1817
14	82.2	1804
15	65.0	1744
16	72.6	1773
17	65.0	1674
18	65.4	1791
19	64.8	1785
20	63.0	1839
21	51.6	1782
22	68.0	1778
23	68.4	1809
24	67.8	1777
25	60.8	1686
26	60.8	1664
27	64.8	1677
28	59.4	1677

データに日本語がある場合

- CSV
- JASPはアウト



データに日本語がある場合

- UTF-8
- ほぼ問題なし

The screenshot shows an Excel spreadsheet with the following data:

No	予防接種	発症の有無
1	無し	無し
2	有り	無し
3	有り	無し
4	無し	無し
5	有り	有り
6	有り	有り
7	無し	有り
8	無し	無し
9	有り	無し
10	有り	無し
11	無し	有り
12	無し	有り
13	有り	無し
14	有り	無し
15	有り	有り
16	有り	無し
17	有り	無し
18	無し	有り
19	有り	無し
20	無し	有り
21	有り	無し
22	無し	無し
23	有り	有り
24	有り	有り
25	無し	有り
26	有り	無し
27	有り	無し
28	有り	無し
29	無し	有り
30	無し	有り

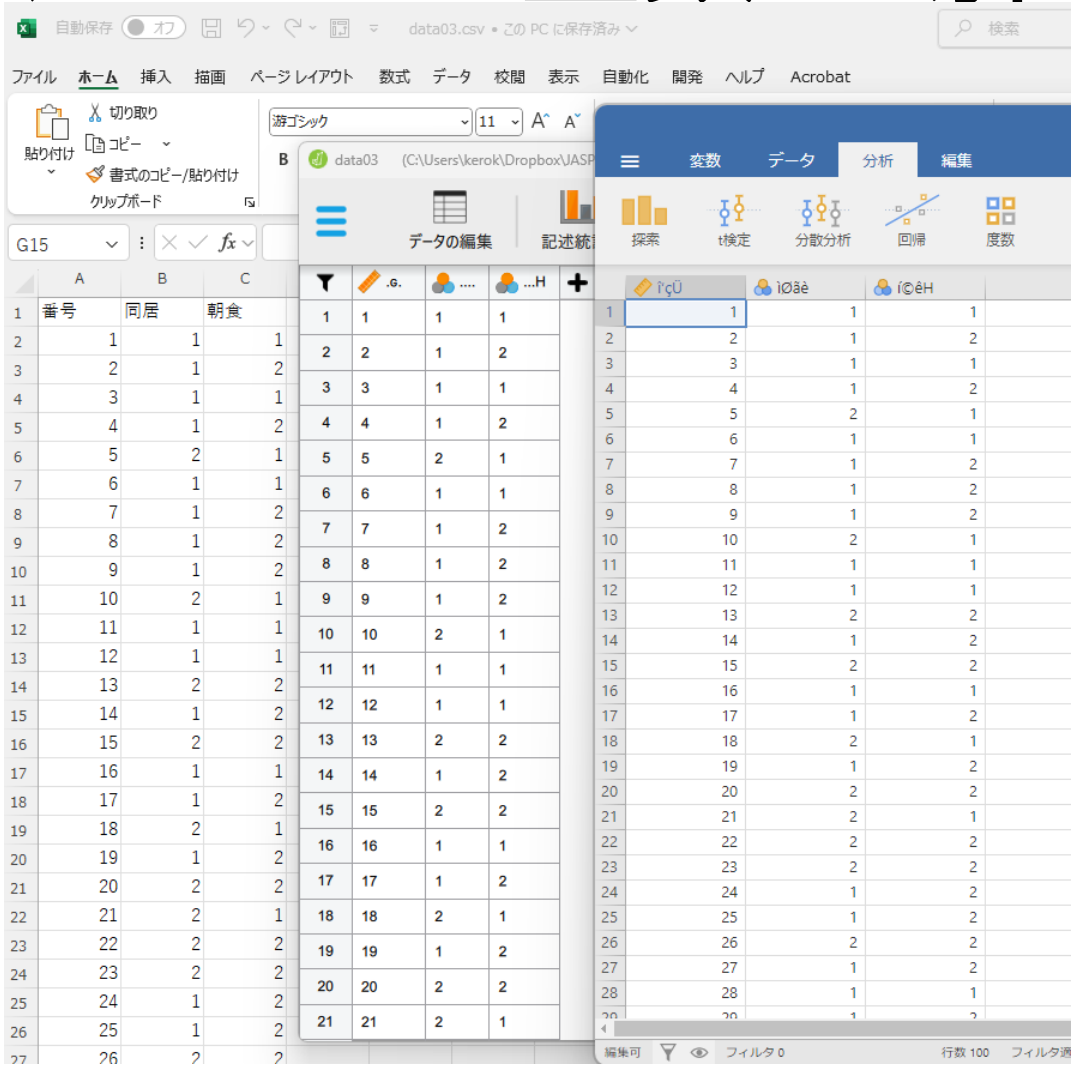
データに日本語がある場合

- Excel
- これは問題なし

	No	予防接種	発症の有無
1	1	無し	無し
2	2	有り	無し
3	3	有り	無し
4	4	無し	無し
5	5	有り	有り
6	6	有り	有り
7	7	無し	有り
8	8	有り	有り
9	9	有り	無し
10	10	有り	無し
11	11	無し	有り
12	12	無し	有り
13	13	無し	有り
14	14	有り	有り
15	15	有り	無し
16	16	有り	無し
17	17	有り	無し
18	18	無し	有り
19	19	有り	無し
20	20	無し	有り
21	21	有り	無し

タイトルだけ日本語で、データが整数の場合

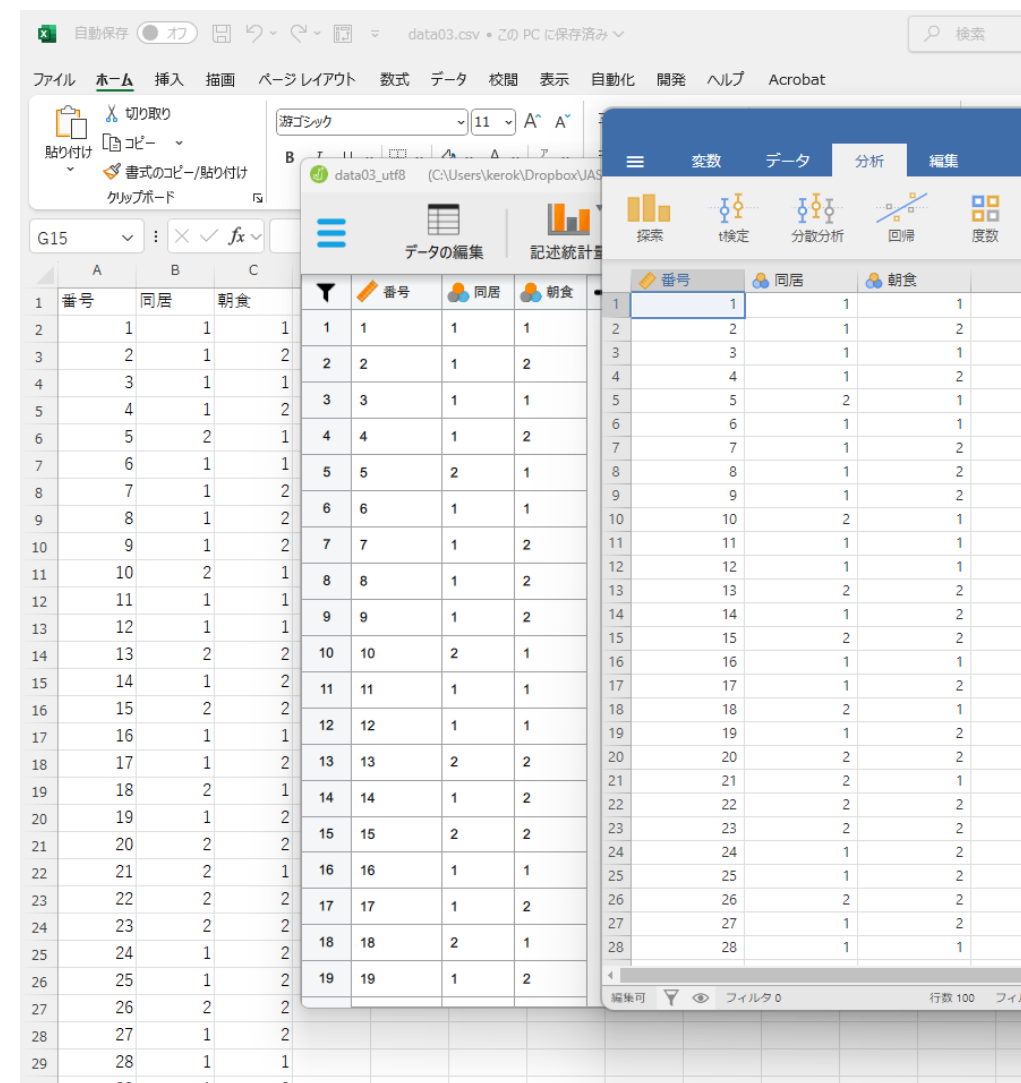
- CSV
- タイトルがダメ



	A	B	C
1	番号	同居	朝食
2	1	1	1
3	2	1	2
4	3	1	1
5	4	1	2
6	5	2	1
7	6	1	1
8	7	1	2
9	8	1	2
10	9	1	2
11	10	2	1
12	11	1	1
13	12	1	1
14	13	2	2
15	14	1	2
16	15	2	2
17	16	1	1
18	17	1	2
19	18	2	1
20	19	1	2
21	20	2	2
22	21	2	1
23	22	2	2
24	23	2	2
25	24	1	2
26	25	1	2
27	26	2	2

タイトルだけ日本語で、データが整数の場合

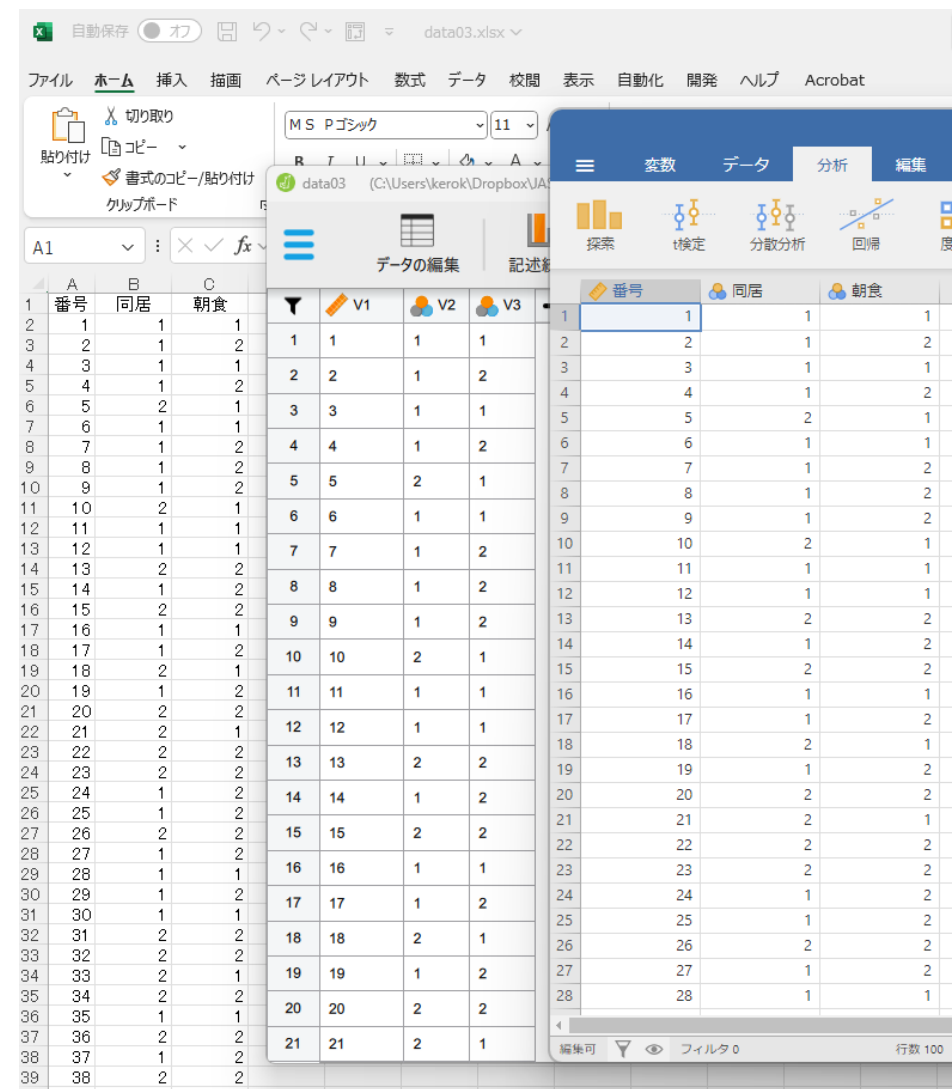
- UTF-8
- 問題なし



	番号	同居	朝食
1	1	1	1
2	2	1	2
3	3	1	1
4	4	1	2
5	5	2	1
6	6	1	1
7	7	1	2
8	8	1	2
9	9	1	2
10	10	2	1
11	11	1	1
12	12	1	1
13	13	2	2
14	14	1	2
15	15	2	2
16	16	1	1
17	17	1	2
18	18	2	1
19	19	1	2
20	20	2	2
21	21	2	1
22	22	2	2
23	23	2	2
24	24	1	2
25	25	1	2
26	26	2	2
27	27	1	2
28	28	1	1

タイトルだけ日本語で、データが整数の場合

- Excel
- JASPはタイトルが置き換わる



The screenshot shows a Microsoft Excel spreadsheet titled 'data03.xlsx'. The worksheet contains a data table with the following structure:

	A	B	C
	番号	同居	朝食
1	1	1	1
2	2	1	2
3	3	1	1
4	4	1	2
5	5	2	1
6	6	1	1
7	7	1	2
8	8	1	2
9	9	1	2
10	10	2	1
11	11	1	1
12	12	1	1
13	13	2	2
14	14	1	2
15	15	2	2
16	16	2	2
17	17	1	1
18	18	1	2
19	19	2	1
20	20	1	2
21	21	2	2
22	22	2	1
23	23	2	2
24	24	2	2
25	25	1	2
26	26	1	2
27	27	2	2
28	28	1	2
29	29	1	1
30	30	1	2
31	31	1	1
32	32	2	2
33	33	2	2
34	34	2	1
35	35	2	2
36	36	1	1
37	37	2	2
38	38	1	2
39	39	2	2

データ入力形式

- ほぼ完璧なのはcsvのUTF-8
 - 普段はExcel形式で管理
 - JASPに渡すときにCSV(UTF-8)にする
 - CSVファイルは複数シートを持ってないので注意
- 他のアプリケーションを使う場合も問題ない
 - jamoviに渡しても問題ない
- Mac版のExcelはUTF-8が原則だけど、Windows版はそうでもないので微妙なときがある

ムダ知識

- アプリがデータファイルを読み込むとき
 - CSVファイルはテキストファイルなので、文字コード情報は原則持っていない
 - BOM (Byte Order Mark)を振ってあって、EF BB BFならUTF-8、FF FE (FE FF)ならUTF-16と判断するアプリもある
 - 全体から「文字コードはこれかな？」と推測している
 - 1行目はデータ名として判断する
 - 1行1列目から始まっていないとコケる
 - セル結合があるとコケる
 - 空白が入っていると空欄の変数を作る

ムダ知識

- アプリがデータファイルを読み込むとき
 - 2行目以降からデータ本体と判断する
 - 2行目のデータを見て、その変数が文字なのか、数値なのか判断する。文字と整数なら名義か順序、小数点データならスケール(連続量)と判断する
 - いきなり欠損値だと、うまく行かないアプリがある
 - JASPは「列の種類」(名義、順序、スケール)がデータ情報
jamoviは「測定値型」(名義、順序、連続、ID)と「データ型」(整数、小数、文字)の2種類で情報を持つ

名義データをどう入力するか

- 文字でそのまま入力の方がわかりやすい
- 数値で入れた方がトラブルは少ない
 - 日本語でもExcelファイルの内容を結構渡してくれる
 - トラブルが起きる原因のほとんどはこれ

個人的な懸念

- アプリでの日本語の受け渡しはトラブルが多い
 - 昔は日本語が使えなかった
 - 文字コードセットの設定が複数ある(Shift-JIS UTF-8など)
 - 海外のソフトは、日本語の扱いに無頓着
 - 日本語話者が少ないから
 - 学術論文は英語がデファクトスタンダード

結論

- データ形式はCSVのUTF-8形式が最強
 - Webでも、WindowsでもMacでも大丈夫
- 名義データは、日本語で入れてもまあ大丈夫かな
 - 古いデータを持ってくる場合にトラブルになることがある
 - 古いテキストデータは、BOMが無いことが多いので、文字コード判定に難があることがある

面倒くさいけど便利かもしれない

データラベル

ラベル

- 名義変数のデータを数値で入れると、その数値に対して表示上の文字を設定できる
- データ自体は元の数値のままなので、並び順は数字になる

ラベル

- ラベルをつけると保存はJASP形式(*.jasp)
 - CSVではなくなる

data02_utf8_label* (C:\Users\kerok\Dropbox\JASP)

データの編集

記述統計量

t検定

分散分析

混合モデル

名前: 予防接種

長い名前: 予防接種

列の種類: 名義

説明: ...

計算された種類: 計算されていない

ラベルエディタ

欠測値

フィルター

値

ラベル

✓	0	無し
✓	1	有り

Add new level: 値 ラベル

▼

No

予防接種

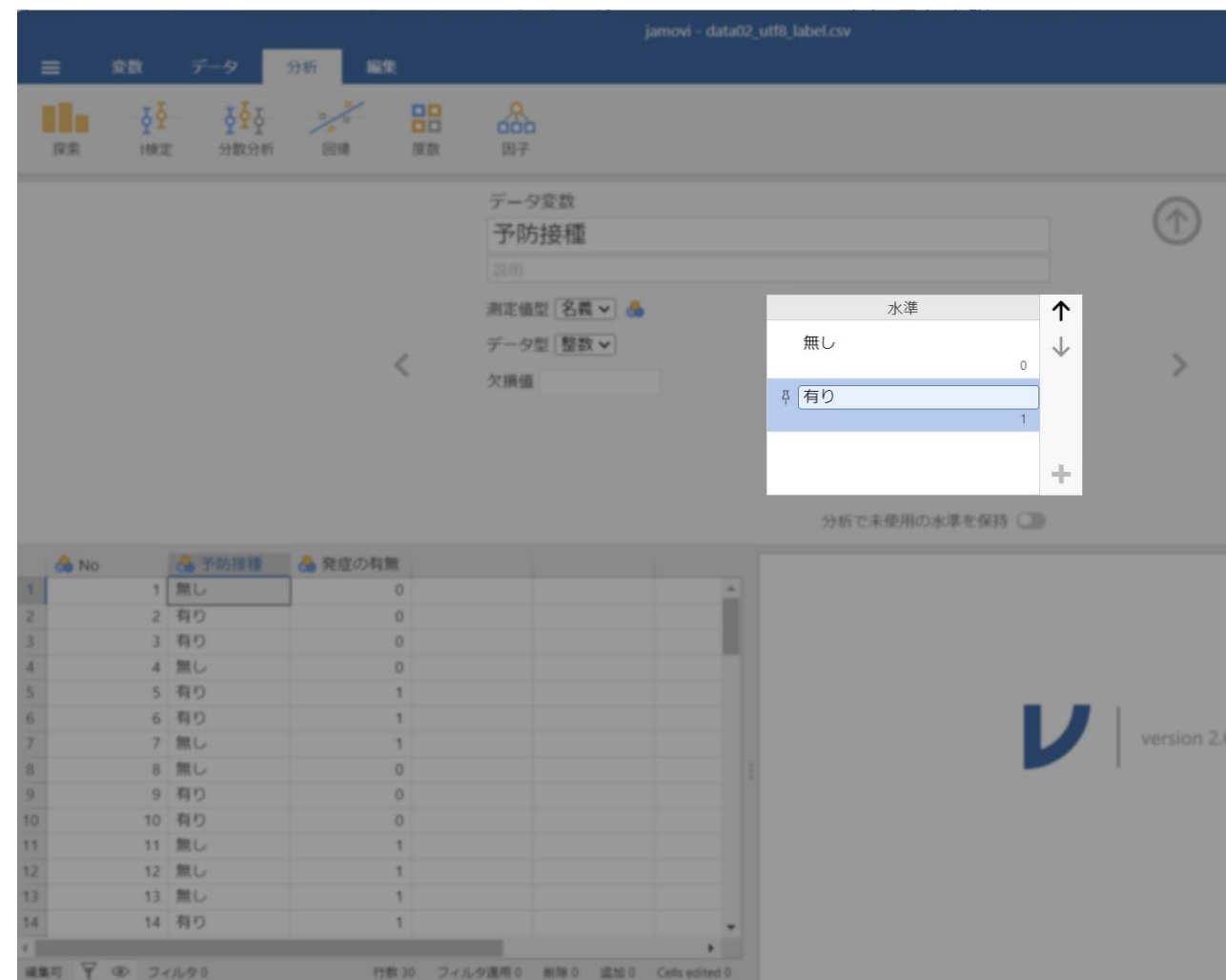
発症の有無

+

1	1	無し	0	0
2	2	有り	1	0
3	3	有り	1	0
4	4	無し	0	0
5	5	有り	1	1
6	6	有り	1	1
7	7	無し	0	1

ラベル

- jamoviでもできる
- jamoviファイル(*.omv)



ラベルの意義

- 文字データでも一応大丈夫なのに、わざわざ数値にしてラベルをつけるのはなぜか
 - 日本語でのデータ入力を信用していないから
 - ラベルにすると、表示したい順番をコントロールできるから

ラベルの意義

- 表示順のコントロール
 - 文字のデータ(カテゴリ)だと、出力は文字コード順
 - それが、自分の望んだ順番とは限らない
 - 数値で入れてラベルにすると、数値の順で表示となる
 - 表示を望んだ順にできる

データラベル

- 左 文字として入力 右 数値データにラベル

data03_1

(C:\Users\kerok\Dropbox\JASP)

データの編集

記述統計量

	番号	居住	朝食
1	1	独居	食べる
2	2	独居	食べない
3	3	独居	食べる
4	4	独居	食べない
5	5	同居	食べる
6	6	独居	食べる
7	7	独居	食べない
8	8	独居	食べない
9	9	独居	食べない
10	10	同居	食べる

data03_utf8

(C:\Users\kerok\Dropbox\JASP)

分析

同期

データのサイ

	番号	同居	朝食
1	1	独居 1	食べる 1
2	2	独居 1	食べない 2
3	3	独居 1	食べる 1
4	4	独居 1	食べない 2
5	5	同居 2	食べる 1
6	6	独居 1	食べる 1
7	7	独居 1	食べない 2
8	8	独居 1	食べない 2
9	9	独居 1	食べない 2
10	10	同居 2	食べる 1

集計すると

- ・ 左 文字として入力 右 数値データにラベル

度数分布表 ▼

居住の頻度 ▼

居住	頻度	パーセント	有効パーセント	累積パーセント
同居	40	40.0	40.0	40.0
独居	60	60.0	60.0	100.0
欠損値	0	0.0		
合計	100	100.0		

朝食の頻度

朝食	頻度	パーセント	有効パーセント	累積パーセント
食べない	50	50.0	50.0	50.0
食べる	50	50.0	50.0	100.0
欠損値	0	0.0		
合計	100	100.0		

度数分布表

同居の頻度

同居	頻度	パーセント	有効パーセント	累積パーセント
独居	60	60.0	60.0	60.0
同居	40	40.0	40.0	100.0
欠損値	0	0.0		
合計	100	100.0		

朝食の頻度

朝食	頻度	パーセント	有効パーセント	累積パーセント
食べる	50	50.0	50.0	50.0
食べない	50	50.0	50.0	100.0
欠損値	0	0.0		
合計	100	100.0		

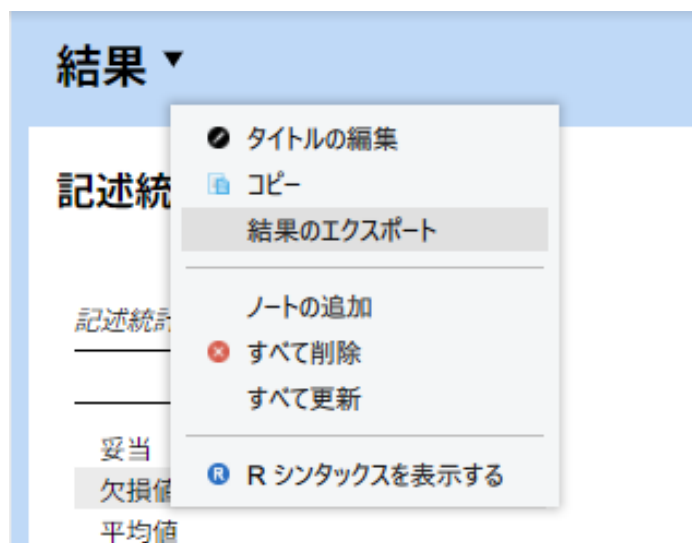
出力の扱い

JASPの出力

- コピペ可能
コピー時はHTMLが原則
TeXも可能
- ファイル出力は
HTML
PDF
- グラフは原則pngかtiff

出力の利用

- 全体
 - HTMLかPDF



ファイルの種類(T):

HTMLファイル (*.html)

HTMLファイル (*.html)

ポータブルドキュメントフォーマット (*.pdf)

出力の利用

・ 部分

・ 個別

度数分布表 ▼

同居の頻度

同居	頻度	パーセント	有効パーセント	累積パーセント
独居	0	0.0	60.0	60.0
同居	40	40.0	40.0	100.0
欠損値	0	0.0		
合計	100	100.0		

朝食の頻度

朝食	頻度	パーセント	有効パーセント	累積パーセント
食べる	50	50.0	50.0	50.0
食べない	50	50.0	50.0	100.0
欠損値	0	0.0		
合計	100	100.0		

同居の頻度 ▼

同居	頻度	パーセント	有効パーセント	累積パーセント
独居	0	0.0	60.0	60.0
同居	40	40.0	40.0	100.0
欠損値	0	0.0		
合計	100	100.0		

コピーをするとHTML形式

- Wordへの貼り付け

度数分布表

同居 の頻度				
同居	頻度	パーセント	有効パーセント	累積パーセント
独居	60	60.0	60.0	60.0
同居	40	40.0	40.0	100.0
欠損値	0	0.0		
合計	100	100.0		

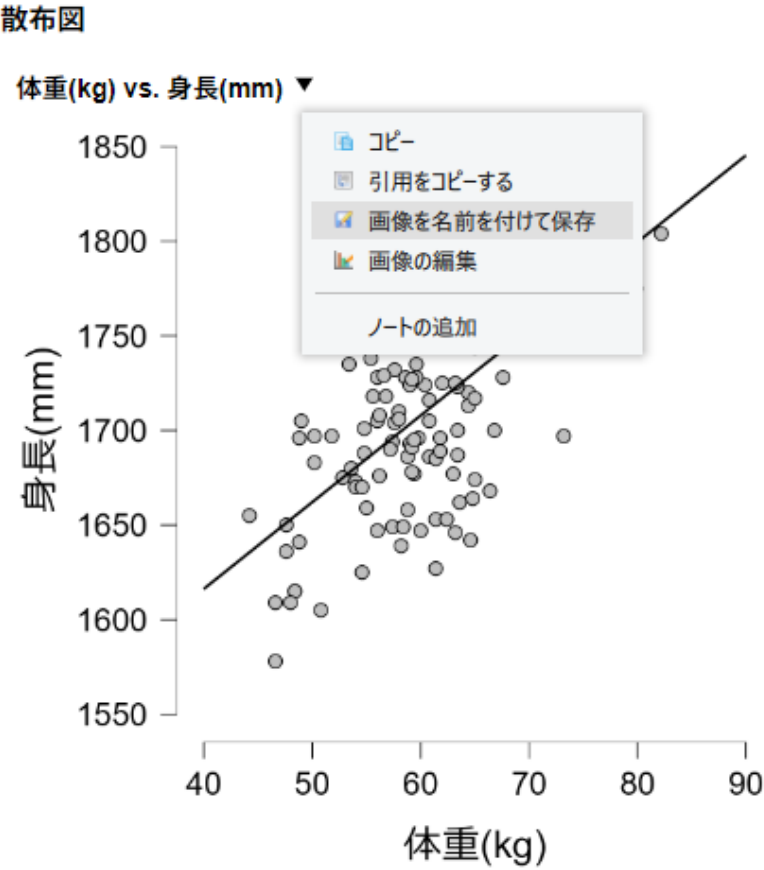
朝食 の頻度

朝食	頻度	パーセント	有効パーセント	累積パーセント
食べる	50	50.0	50.0	50.0
食べない	50	50.0	50.0	100.0
欠損値	0	0.0		
合計	100	100.0		

Ctrl

グラフ

- 形式

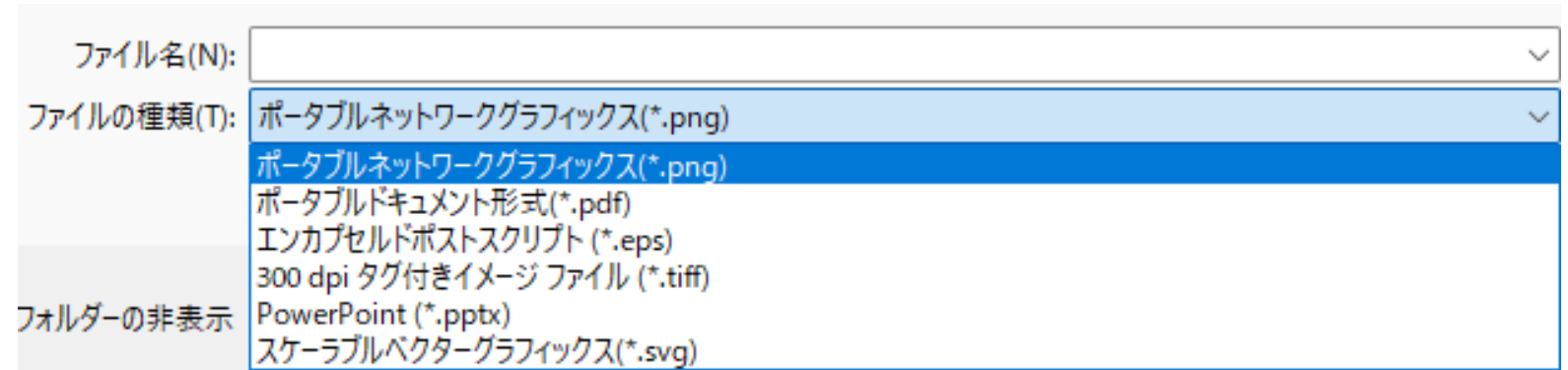


ファイル名(N):	<input type="text"/>
ファイルの種類(T):	<div>ポータブルネットワークグラフィックス(*.png) ポータブルネットワークグラフィックス(*.png) ポータブルドキュメント形式(*.pdf) エンカプセルドポストスクリプト(*.eps) 300 dpi タグ付きイメージ ファイル(*.tiff) PowerPoint(*.pptx) スケーラブルベクターグラフィックス(*.svg)</div>
フォルダーの非表示	

※全体の一括保存を選択したときにグラフがあると、png
ファイルとなって、HTML内で埋め込まれる

グラフ出力形式

- png
- tiff
- pdf
- pptx(PowerPoint)



まとめ

- 入力データはExcelでCSVのUTF-8形式で作成する
といい
- 名義データは数値で入力して、JASPでラベルをつけた
方がいい
- 出力は、場合によってはExcel等で編集した方がいい
 - 棒グラフや円グラフ、ヒストグラム等の基本グラフは、Excel
の方が編集しやすくもいいかもしれない