「神エクセル」自動整形ツール試作版

長野大学企業情報学部　前川ゼミ

シュク　メイスイ

2019/01/11 @GEEKLAB.NAGANO

そもそも「神エクセル」とは？

- 政府統計の総合窓口や自治体などで公開されているエクセル仕様のデータなのに…
  - 機械可読性が低い
    - 「データ抽出禁止の設定がなされていたり画像化されていたりして、集約や再解析にたいへん手間がかかった」
  - （入力の段階でデータの再利用を考えず）罫線を多用した紙の帳票作成を最終目的とするものが多い
  - 造語「神エクセル」
    - 神＝紙
    - 「紙（への出力しか考えていない）エクセル」の意味も含んでいる
- 上記の内容は三重大学の奥村晴彦教授の論文から引用した

奥村氏が挙げた例①

	A	B	C	D	E	F	G
1	[基本集計]				長期時系列表１　（１）労働力人口 － 全国，月別結果		
2	[Basic Tabulation]				Historical data 1　（1） Labour force - Whole Japan, Monthly Data		
3							
4					（万人）	(Ten thousand persons)	
5					原数値（2010年国勢調査基準切り替え以前の既公表値）		
6					Original series (initially released data before 2010-Census base revision)		
7					男女計	男	女
8	年　月						
9	Year and month				Both sexes	Male	Female
10							
12	平成17年	10月	Oct.		6713	3930	2783
13	2005	11月	Nov.		6636	3901	2736
14		12月	Dec.		6580	3881	2699
15	平成18年	1月	Jan.		6561	3864	2697
16	2006	2月	Feb.		6549	3855	2694
17		3月	Mar.		6597	3887	2710
18		4月	Apr.		6652	3901	2751
19		5月	May		6725	3925	2799
20		6月	June		6717	3924	2793
21		7月	July		6688	3908	2780
22		8月	Aug.		6699	3912	2788
23		9月	Sept.		6711	3917	2794
24		10月	Oct.		6718	3921	2797
25		11月	Nov.		6669	3892	2777
26		12月	Dec.		6598	3873	2725
27	平成19年	1月	Jan.		6542	3861	2681
28	2007	2月	Feb.		6572	3878	2694
29		3月	Mar.		6632	3898	2733
30		4月	Apr.		6712	3925	2788
31		5月	May		6757	3942	2814

暗にセルを結合した表

奥村氏が挙げた例②

平成25年４月１日現在　住民基本台帳による年齢別、男女別人口

◇5歳階級別、男女別人口　前年比

平成25年４月１日				平成24年４月１日		
計	男	女	年齢	計	男	女
8,423	4,268	4,155	0～4	8,285	4,289	3,996
7,771	3,984	3,787	5～9	7,602	3,843	3,759
7,260	3,652	3,608	10～14	7,135	3,603	3,532
7,114	3,678	3,436	15～19	6,947	3,566	3,381
7,584	3,933	3,651	20～24	7,722	4,026	3,696
9,875	4,997	4,878	25～29	10,078	5,096	4,982
12,685	6,367	6,318	30～34	12,698	6,466	6,232
14,569	7,570	6,999	35～39	14,629	7,630	6,999
14,033	7,337	6,696	40～44	13,146	6,934	6,212
10,516	5,378	5,138	45～49	9,873	5,050	4,823
9,072	4,613	4,459	50～54	8,701	4,398	4,303
9,118	4,357	4,761	55～59	9,561	4,542	5,019
12,399	5,903	6,496	60～64	13,330	6,381	6,949
12,055	5,768	6,287	65～69	11,320	5,465	5,855
10,078	4,968	5,110	70～74	9,519	4,748	4,771
7,114	3,435	3,679	75～79	6,818	3,250	3,568
4,515	1,926	2,589	80～84	4,174	1,776	2,398
2,437	881	1,556	85～89	2,290	793	1,497
1,011	217	794	90～94	987	237	750
347	73	274	95～99	334	66	268
48	6	42	100～	46	4	42
168,024	83,311	84,713	総数	165,195	82,163	83,032

◇各歳別、男女別人口

年齢	計	男	女	年齢	計	男	女
0	1,647	792	855	50	1,947	1,007	940
1	1,702	875	827	51	1,887	966	921
2	1,727	883	844	52	1,710	870	840
3	1,707	887	820	53	1,784	907	877
4	1,640	831	809	54	1,744	863	881
5	1,613	853	760	55	1,679	814	865
6	1,573	830	743	56	1,699	826	873
7	1,514	776	738	57	1,834	884	950
8	1,522	768	754	58	1,978	923	1,055
9	1,549	757	792	59	1,928	910	1,018
10	1,481	748	733	60	2,130	990	1,140
11	1,443	731	712	61	2,318	1,114	1,204
12	1,487	754	733	62	2,416	1,159	1,257
13	1,365	702	663	63	2,701	1,286	1,415
14	1,484	717	767	64	2,834	1,354	1,480
15	1,422	725	697	65	3,047	1,455	1,592
16	1,372	706	666	66	2,475	1,182	1,293
17	1,424	709	715	67	1,761	856	905
18	1,505	794	711	68	2,221	1,052	1,169
19	1,391	744	647	69	2,551	1,223	1,328
20	1,382	693	689	70	2,244	1,105	1,139
21	1,540	820	720	71	2,350	1,159	1,191
22	1,474	779	695	72	2,070	1,003	1,067
23	1,569	794	775	73	1,883	946	937
24	1,619	847	772	74	1,531	755	776
25	1,667	868	799	75	1,558	787	771

二つの表を一つのシートに入力し，しかも2番目の表が長いので二つに分割した例
流山市　http://www.city.nagareyama.chiba.jp/10763/011144.html

奥村氏が挙げた例③

※VDT作業（パソコン等を使用し，データ入力，検索，文書等作成，プログラミング等を行う作業）についてお尋ねします。

フリガナ		性別	生年月日	部局名	
氏名		男・女	年　月　日（　才）	職名	

職務内容（該当職務に○をつける）	1. 医療事務	2. 医療技術	3. 図書業務	4. 情報処理・技術
	5. 一般事務	6. 教員	7. その他（　）	

VDT作業の有無	有　・　無	「有」の方のみ以下の質問にお答えください。	

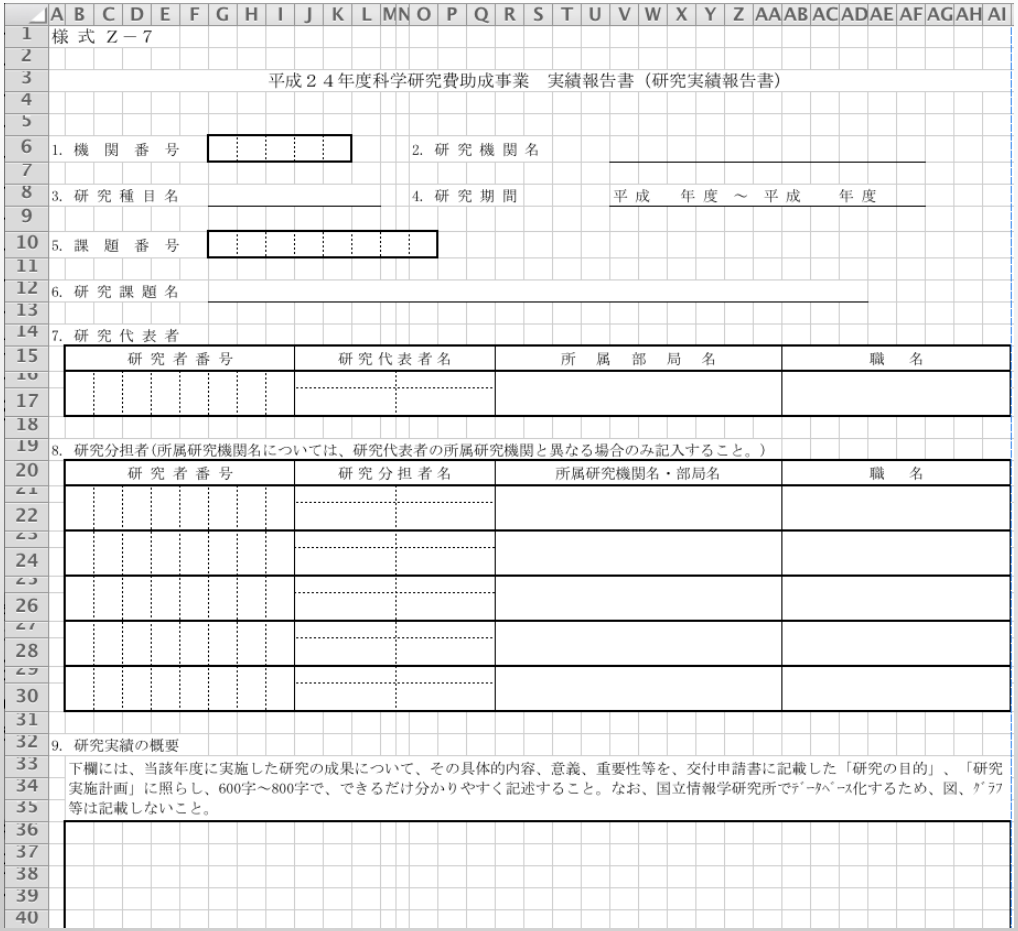
病歴	眼圧が高いと診断されたことがありますか			VDT作業歴	年　月　～　年　月
	有	（　年頃）	無		
	緑内障と診断されたことがありますか				
	有	（　年頃）	無		

VDT作業の状況	（注2）VDT作業区分（該当区分に○をつける）	A		（注2）作業の種類（該当の種類に○をつける）	1. 単純入力型
		B			2. 拘束型
		C			3. 対話型
	1回の平均VDT従事時間数	時間　分			4. 技術型
	1日の平均VDT従事時間数	時間　分			5. 監視型
	1週間の平均VDT従事日数	日			6. その他の型

	項目	有無・適否		項目	有無・適否	
自覚症状	目が疲れる	有	無	頭が重い（痛い）	有	無
	目が痛い	有	無	肩や頸がこる	有	無
	目がかすむ	有	無	手指の力が入りにくい	有	無
	目が乾く	有	無	手指がしびれる（痛い）	有	無
	ものがちらついて見える	有	無	腕がだるい（痛い）	有	無

「図形」で○を付けることを想定したExcelアンケート

奥村氏が挙げた例④



様 式 Ｚ－７

平成２４年度科学研究費助成事業　実績報告書（研究実績報告書）

1. 機 関 番 号　　2. 研 究 機 関 名

3. 研 究 種 目 名　　4. 研 究 期 間　　平成　年度 ～ 平成　年度

5. 課 題 番 号

6. 研 究 課 題 名

7. 研 究 代 表 者

研究者番号	研究代表者名	所属部局名	職名

8. 研究分担者（所属研究機関名については、研究代表者の所属研究機関と異なる場合のみ記入すること。）

研究者番号	研究分担者名	所属研究機関名・部局名	職名

9. 研究実績の概要

下欄には、当該年度に実施した研究の成果について、その具体的内容、意義、重要性等を、交付申請書に記載した「研究の目的」、「研究実施計画」に照らし、600字～800字で、できるだけ分かりやすく記述すること。なお、国立情報学研究所でﾃﾞｰﾀﾍﾞｰｽ化するため、図、ｸﾞﾗﾌ等は記載しないこと。

「平成24年度科学研究費助成事業実績報告書（研究実績報告書）」
https://www.jsps.go.jp/j-grantsinaid/16_rule/index.html 様式Z-7

機械可読性とは？①

- パソコンが自動的にデータを読み取る能力
- 読み取ったデータは{キー：値}の格式で保存される
  - {地域：[上田, 東御, 長和], 人口数: [100万, 80万, 70万]}
- こうしたデータは「データベース化」されたデータ
- 自動的にデータを表（エクセル、CSV…）から読み取るのが「論理」は必要
  - 表のどこからどこまではデータ、どこは見出し行、どこはコメント
  - 以上のことを判断する
  - 「数字はデータ、数字でなければ見出し」
  - 「人、世帯数を書いている『セル』は単位」

機械可読性とは？②

- 一番論理を手軽に作成できる表の構造：
  二次元構造
  - 表の天然の構造
  - 1行1列の見出しと、それに囲まれたデータ
  - 列ごとにデータを読み取れば論理になる
  - {年度：[H28, H29, H30…], 世代数：[10万, 15万, 13万…], 男性人口数：[5万, 7万, 6万…]}
- 想像の中にしか存在していない構造
- 現実な表の構造（エクセル）は？

上田市のデータ①

1 人口及び世帯数（各年10月1日現在）

		平成17年	平成22年	平成27年
人口（人）	総数	163,651	159,597	156,827
	対前回増減数	△ 2,917	△ 4,054	△ 2,770
	対前回増減率	△ 1.8%	△ 2.5%	△ 1.7%
	男	79,770	77,589	76,776
	対前回増減数	△ 2,041	△ 2,181	△ 813
	対前回増減率	△ 2.5%	△ 2.7%	△ 1.0%
	女	83,881	82,008	80,051
	対前回増減数	△ 876	△ 1,873	△ 1,957
	対前回増減率	△ 1.0%	△ 2.2%	△ 2.4%
世帯数（世帯）		59,858	60,660	62,696
	対前回増減数	339	802	2,036
	対前回増減率	0.6%	1.3%	3.4%

上田市のデータ②

9　男女別5歳階級別人口の推移（新上田市）

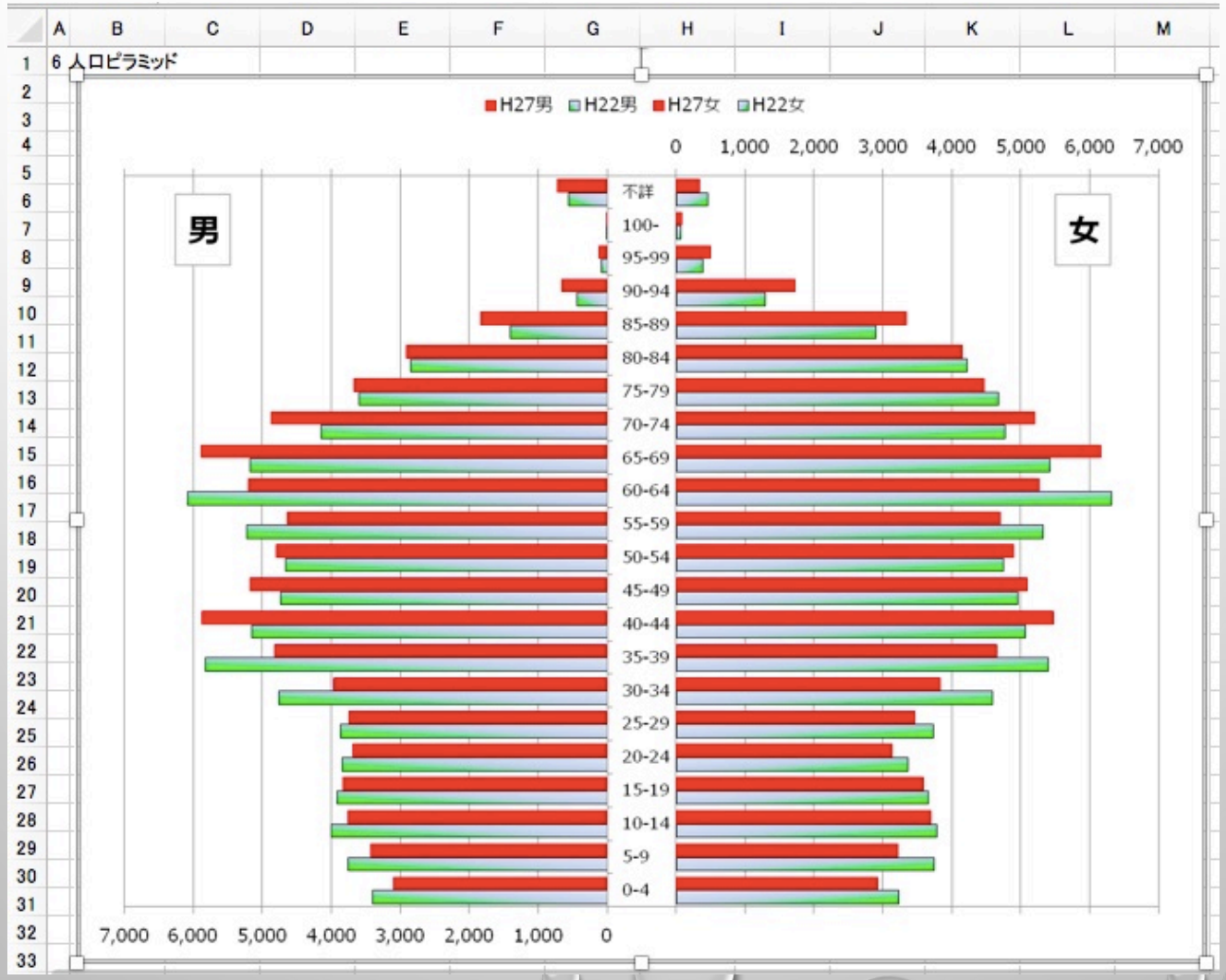
年齢区分	昭和35年			昭和40年		
	計	男	女	計	男	女
合計	138,170	65,373	72,797	138,001	65,132	72,869
0～4	10,421	5,325	5,096	10,527	5,388	5,139
5～9	12,201	6,182	6,019	10,450	5,318	5,132
10～14	15,425	7,790	7,635	12,204	6,123	6,081
15～19	13,959	6,605	7,354	14,455	6,744	7,711
20～24	10,317	4,547	5,770	10,730	4,596	6,134
25～29	10,479	4,817	5,662	9,629	4,513	5,116

上田市のデータ③

8　国勢調査年次別の世帯数及び人口

地区名			大正9年10月1日現在			
			世帯数	人口		
				計	男	女
				人	人	人
総数			22,358	112,001	52,896	59,105
上田	計		15,093	71,928	34,734	37,194
	上田	計	9,885	46,533	22,427	24,106
		上田 計	5,656	26,271	12,588	13,683
		上田 東部				
		上田 南部				
		上田 中央				
		上田 北部				
		上田 西部				
		城下	817	3,681	1,806	1,875
		塩尻	619	2,959	1,402	1,557
		川辺	590	2,851	1,410	1,441
		泉田	360	1,615	803	812
		神科	1,103	5,336	2,584	2,752
		神川	740	3,820	1,834	1,986
	豊殿	計	778	3,964	1,939	2,025
		豊里	391	2,050	1,001	1,049
		殿城	387	1,914	938	976
	塩田	計	3,100	14,763	7,180	7,583
		東塩田	1,052	5,039	2,435	2,604
		中塩田	970	4,758	2,304	2,454
		西塩田	733	3,462	1,711	1,751
		別所	345	1,504	730	774
	川西	計	1,330	6,668	3,188	3,480
		浦里	663	3,459	1,588	1,871
		室賀	461	2,125	1,078	1,047
		小泉	206	1,084	522	562

上田市のデータ④


6 人口ピラミッド
H27男
H22男
H27女
H22女
0
1,000
2,000
3,000
4,000
5,000
6,000
7,000
男
女
不詳
100-
95-99
90-94
85-89
80-84
75-79
70-74
65-69
60-64
55-59
50-54
45-49
40-44
35-39
30-34
25-29
20-24
15-19
10-14
5-9
0-4
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0

CSV仕様①

- パソコンが最も扱いやすい表データの仕様
  - 「具体的には、CSVやXMLを推奨するが…」（内閣官房IT担当室「数値（表）、文章、地理空間情報のデータ作成に当たっての留意事項（案）」）
- パソコンがデータをデータベースに保存するように、文字・数字以外のもの（写真、図表など）を扱えない
- コンマ「,」でセルを区切り、改行で行を区切る
- ただのテキストファイル

CSV仕様②

- 地区, 世代数, 男性人口数, 女性人口数
- 上田, 12500, 25000, 25000

地区	世代数	男性人口数	女性人口数
上田	12500	25000	25000

- 先ほどの例を全部そのままCSVにして？

上田市のデータ①

1 人口及び世帯数（各年10月1日現在）					
			平成17年	平成22年	平成27年
	人口（人）	総数	163,651	159,597	156,827
		対前回増減数	△ 2,917	△ 4,054	△ 2,770
		対前回増減率	△ 1.8%	△ 2.5%	△ 1.7%
		男	79,770	77,589	76,776
		対前回増減数	△ 2,041	△ 2,181	△ 813
		対前回増減率	△ 2.5%	△ 2.7%	△ 1.0%
		女	83,881	82,008	80,051
		対前回増減数	△ 876	△ 1,873	△ 1,957
		対前回増減率	△ 1.0%	△ 2.2%	△ 2.4%
	世帯数（世帯）		59,858	60,660	62,696
		対前回増減数	339	802	2,036
		対前回増減率	0.60%	1.30%	3.40%

上田市のデータ②

	9　男女別５歳階級別人口の推移（新上田市）					
	昭和35年			昭和40年		
年齢区分	計	男	女	計	男	女
合計	138,170	65,373	72,797	138,001	65,132	72,869
0～4	10,421	5,325	5,096	10,527	5,388	5,139
5～9	12,201	6,182	6,019	10,450	5,318	5,132
10～14	15,425	7,790	7,635	12,204	6,123	6,081
15～19	13,959	6,605	7,354	14,455	6,744	7,711
20～24	10,317	4,547	5,770	10,730	4,596	6,134
25～29	10,479	4,817	5,662	9,629	4,513	5,116

上田市のデータ③

8　国勢調査年次別の世帯数及び人口								
	地区名				大正9年10月1日現在			
					世帯数	人口		
						計	男	女
						人	人	人
	総数				22,358	112,001	52,896	59,105
	上田	計			15,093	71,928	34,734	37,194
		上田		計	9,885	46,533	22,427	24,106
			上田	計	5,656	26,271	12,588	13,683
				東部				
				南部				
				中央				
				北部				
				西部				
				城下	817	3,681	1,806	1,875
				塩尻	619	2,959	1,402	1,557
				川辺	590	2,851	1,410	1,441
				泉田	360	1,615	803	812
				神科	1,103	5,336	2,584	2,752
				神川	740	3,820	1,834	1,986
		豊殿		計	778	3,964	1,939	2,025
				豊里	391	2,050	1,001	1,049
				殿城	387	1,914	938	976

上田市のデータ④

6 人口ピラミッド

CSV化された結果

- CSV化された＝一切の美化や印刷用の準備を抜きにした
- CSV化された＝「パソコンが扱うデータ」化
  - パソコンでは「セル」と「セルの座標」しか分からない
- いくつかのパターンが出た：
  - 見出し行は1行、見出し列は2列以上
  - 見出し列は1列、見出し行は2行以上
  - 見出し行・列はともに複数
  - データが写真や図表などのため見出しどころか
    タイトルしか残っていない
- 上記の大前提は「データは数字」

今度試作したツールについて①

- 対象:「データは数字であり、見出し行・列はともに複数」の神エクセルファイル
  - 例③に基づいて作成した
- 論理:
  - 行・列ごとにセルを読み込み、その中の「数字である」セルの個数を計算し、頻度が一番高い数字がデータの区間になる
  - 行・列のセル個数からデータの区間の長さを引いた結果は見出しの区間

今度試作したツールについて②

- 見出し行・列の特徴：

上田	計		
	上田		計
		上田	計

（見出し列）

行ごとに読み取ると…

- [上田,計,(空白),(空白)]
- [(空白),上田,(空白),計]
- [(空白),(空白), 上田, 計]

今度試作したツールについて③

- セルごとに空白であるかどうかを判断…
- 空白の場合：前行からセルを読み取り、空白セルを充填する…
- 空白でない場合はこの繰り返しから飛び出す…
- 上記の行為を行ごとに繰り返す結果：
  - [上田,計,(空白),(空白)]
  - [上田,上田,(空白),計]
  - [上田,上田, 上田, 計]
- コンマ「,」をアンダーバー「_」に入れ替える：
  - 上田_計_(空白)_(空白)
  - 上田_上田_(空白)_計
  - 上田_上田_上田_計

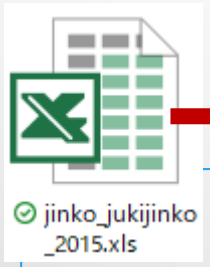
試作したツールについて④

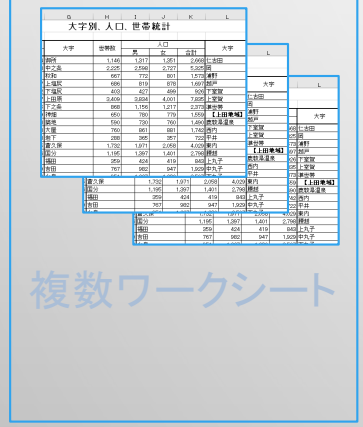
- 複数の見出し行・列を一つにまとめることができる
- 行ごとに見出し行→見出し列の1つ目のセルとデータ区間の1行目のデータを新しいCSVファイルに書き込んで保存

- 他のパターンとの互換度：70%～
  - パターンが増えると論理がいっそう複雑になる
  - 誤差が避けられない
- 現在一番期待していること：
  - きちんとした「1行1列の見出しを持つエクセルファイル」がこれから世の中に広まっていくこと

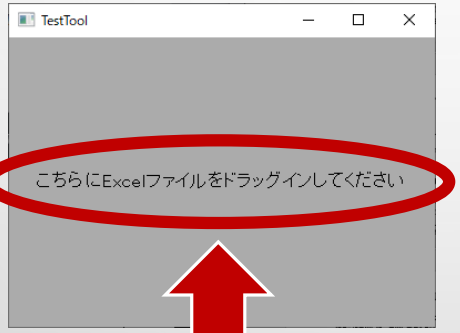
ツールの実演

統計Excelファイル

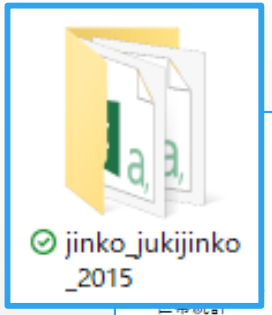




Excel→CSV
一括変換



ここにExcelファイル
をドラッグ＆ドロップ
する



CSVファイル(ワークシート別)

ご静聴ありがとう
ございました