3. 実データによるデータサイエ ンス・AI の演習



アウトライン

- 1. 政府統計データ
- 2. クロス集計表 (Excel を使用)
- 3. 相関 (Excel を使用)
- 4. 平均, 誤差
- 5. オープンデータ

人工知能と統計の関係

調査によって 得られたデータ



数量による 現象の把握 (**統計**)



データを用いた 学習による上達 (**人工知能**)

人工知能が「統計による結論」とは違った振る舞い をするとは考えにくい

3

3-1 政府統計データ

統計

統計(とうけい、statistic)は、**現象を調査**することによって**数量で把握**すること、または、**調査**によって得られた**数量データ(統計量)**のこと

WikiPedia より

総務省統計局が公開する統計表

- 人口・世帯に関する統計
- **企業活動**に関する統計
- **産業構造・波及効果分析**や各種経済統計の基準値となる統計
- **情報通信・科学技術**に関する統計
- **労働・賃金**に関する統計
- ・物価・地価に関する統計
- **住宅・土地**に関する統計
- 家計に関する統計
- 文化に関する統計
- 総合統計書

ダウンロードページ

総務省統計局が公開する統計表

- オンラインで公開
- ・ダウンロードページ

https://www.stat.go.jp/data/guide/download/index.html

人口推計の結果の概要 令和4年4月報より(抜粋)

2022年4月1日現在 (概算値) (令和4年) April 1, 2022 (Provisional estimates) 年 齢 階級 総人口 Total population 男女計 女 Age groups Both sexes Male 人 口(単位 万人) Population (Ten thousand persons) **Total** 12519 6086 0 ~ 4歳 433 221 vears old $5 \sim 9$ 499 256 243 273 $10 \sim 14$ 533 260 $15 \sim 19$ 556 285 270 625 320 $20 \sim 24$ $25 \sim 29$ 638 328 649 332 $30 \sim 34$ 317 729 371 359 $40 \sim 44$ 806 409 397 958 485 934 790 $55 \sim 59$ 395 395 $60 \sim 64$ 740 366 374 $65 \sim 69$ 767 373 394 $70 \sim 74$ $75 \sim 79$ 679 304 80 ~ 84 568 237 332 393 142 251 $85 \sim 89$ 197 55 141 $90 \sim 94$ $95 \sim 99$ 57 11 46 100歳以上 and over

労働力調査(基本集計)2022年 (令和4年)3月分結果より(抜粋)

	年平均			月次(季節調整値)				
	2019年	2020年	2021年	2021年12月	2022年1月	2月	3月	
完全失業率	2.4%	2.8%	2.8%	2.7%	2.8%	2.7%	2.6%	

2020年基準 消費者物価指数 全国 2022 年(令和4年)3月分より(抜粋)

	年平均(前年比%)			月次(前年同月比%)			
	2019年	2020年	2021年	2021年12月	2022年1月	2月	3月
総合	0.5	0.0	▲0.2	0.8	0.5	0.9	1.2
生鮮食品を除く総合	0.6	▲0.2	▲0.2	0.5	0.2	0.6	0.8
生鮮食品及びエネルギーを除く総合	0.6	0.2	▲0.5	▲0.7	▲ 1.1	▲1.0	▲0.7

e-Stat

政府統計(総務省統計局以外も網羅)のポータル サイト



e-Stat での統計データの検索からダウンロード

- ① 統計データの検索
 - ・**分野**による検索



キーワードによる検索,その他



- ② 必用な項目を選ぶ (表示項目選択)
- ③ レイアウトの設定
- ④ ダウンロード手元のパソコンにファイルがダウンロードされる

家計調査

・全国約9千世帯を対象として、**家計の収入・支出、貯蓄・ 負債**などを毎月調査

e Stat 家計調査のデータを用いた分析

「プリン」にお金を使うことが多い都市は?

	Α	В	С	D	Е	F	G	Н	1	J
7	***	調査又は賃	集計してい:	ないもの						
8	-	該当数字左	がないもの							
9	X	数値が秘囲	置されてい	るもの						
10					N					
11					<i>γ</i> ₂				/品目分類	10800070
12	表章項目	表章項目	世帯区分	世帯区分	時間軸(月	時間軸(月	地域区分	地域区分	/品目分類	348 プリン
13	1	金額	3	二人以上の	2.02E+09	#######	0	全国		133
14	1	金額	3	二人以上の	2.02E+09	#######	1003	01100 札巾	晃市	158
15	1	金額	3	二人以上の	2.02E+09	#######	2003	02201 青潮	 森市	96
16	1	金額	3	二人以上の	2.02E+09	#######	3003	03201 盛	岡市	162
17	1	金額	3	二人以上の	2.02E+09	#######	4003	04100 仙市	台市	85
18	1	金額	3	二人以上の	2.02E+09	#######	5003	05201 秋日	田市	99
19	1	金額	3	二人以上の	2.02E+09	#######	6003	06201 山∄	形市	112
20	1	全類	3	- X W F a	2 N2F+N9	#######	7003	07201 福日	急 市	105

数量による把握

謝辞

https://www.youtube.com/watch?v=0RSIrNTFnJw を参考にし、説明の追加等を行っている

e Stat 家計調査データのダウンロード

① 統計データの検索

分野での検索:**企業・家計・経済→家計調査**

キーワード検索:家計調査



② 検索結果から「家計調査」を選ぶ



③ 種類を選ぶ



④ 品目区分(2020年改訂)(総数:金額)を選ぶ

提供周期	月次			
表番号	統計表	調査年月	公開(更新)日	表示・ダウンロ
用途分類				
001	用途分類(総数)	-	2022-04-05	→ DB 🗪 API
002	用途分類(年間収入五分位階級別)	-	2022-04-05	→ DB
003	用途分類(世帯主の定期収入五分位階級別)	-	2022-04-05	→ DB 🗪 API
004	用途分類(世帯人員別)	-	2022-04-05	→ DB 🗪 API
005	用途分類(世帯主の年齢階級別)	-	2022-04-05	→ DB 🗪 API
006	用途分類 (世帯主の職業別)	-	2022-04-05	→ DB 🗪 API
008	用途分類 (世帯主の勤め先企業規模別)	-	2022-04-05	→ DB 🗪 API
009	用途分類(住居の所有関係別)	-	2022-04-05	→ DB 🗪 API
品目分類				
010	品目分類(2020年改定)(総数:金額)	-	2022-04-0	→ DB 🗪 (PI
011	品目分類(2020年改定)(総数:数量)	-	2022-04-05	→ DB 🗪 API
010	品目分類(平成27年改定)(総数:金額)	-	2020-02-07	→ DB 🗪 API
011	品目分類(平成27年改定)(総数:数量)	-	2020-02-07	→ DB
010	品目分類(平成22年改定)(総数:金額)	-	2015-02-06	→ DB 🗪 API
011	品目分類(平成22年改定)(総数:数量)	-	2015-01-30	→ DB 🗪 API
010	品目分類(平成17年改定) (総数:金額)	-	2010-01-29	→ DB 🗪 API
011	品目分類(平成17年改定)(総数:数量)	-	2010-01-29	→ DB 🗪 API

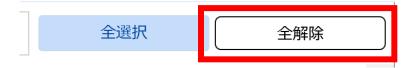
⑤ 「表示項目選択」を展開



⑥「品目分類」の「項目を選択」をクリック



「全解除」をクリック



「プリン」で検索



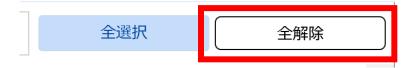
「プリン」にチェック、「確定」をクリック



⑦「時間軸(月次)」の「項目を選択」をクリック



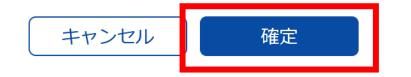
「全解除」をクリック



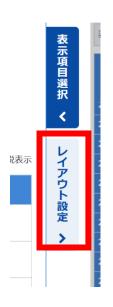
スクロールし、「2022年2月」をチェック



「確定」をクリック



⑧「レイアウト設定」をクリック



⑨ 「地域区分」は「行」のところにドラッグ&ド ロップ. 「設定して表示を更新」



⑩ プリンについての結果を確認

		348 プリン (円)
2022年2月	全国	133 ^
	01100 札幌市	158
	02201 青森市	96
	03201 盛岡市	162
	04100 仙台市	85
	05201 秋田市	99
	06201 山形市	112
	07201 福島市	105
	08201 水戸市	142
	09201 宇都宮市	207
	10201 前橋市	123
	11100 さいたま市	210
	12100 千葉市	147
	13100 東京都区部	105
	14100 横浜市	152
	15100 新潟市	130
	A COOL PULL	

⑪ 「ダウンロード」をクリック

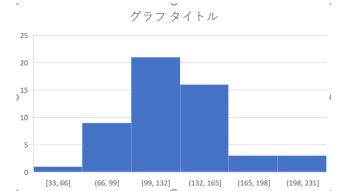


⑪ ダウンロードしたファイルを Excel で開くこと ができる

	А	В	С	D	Е	F	G	Н	I	J	
7	***	調査又は賃	集計してい:	ないもの							
8	-	該当数字が	がないもの								
9	Χ	数値が秘園	匿されてい	るもの							
10					_						
11					h 7				/品目分類	10800070	
12	表章項目	表章項目	世帯区分	世帯区分	時間軸(月	時間軸(月	地域区分	地域区分	/品目分類	348 プリン	/
13	1	金額	3	二人以上の	2.02E+09	#######	0	全国		133	
14	1	金額	3	二人以上の	2.02E+09	#######	1003	01100 札帆	晃市	158	
15	1	金額	3	二人以上の	2.02E+09	#######	2003	02201 青森	集市	96	
16	1	金額	3	二人以上の	2.02E+09	#######	3003	03201 盛岡	岡市	162	
17	1	金額	3	二人以上の	2.02E+09	#######	4003	04100 仙台	台市	85	
18	1	金額	3	二人以上の	2.02E+09	#######	5003	05201 秋日	日市	99	
19	1	金額	3	二人以上の	2.02E+09	#######	6003	06201 山州	形市	112	
20	1	全類	3	- X W F 0	2 N2F+N9	#######	7003	07201 福島	皇 市	105	

⑬ ヒストグラムの作成(範囲を選び、「挿入」でヒ

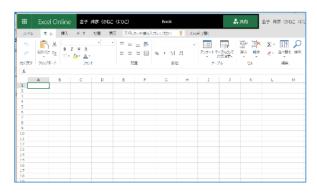
ストグラム)



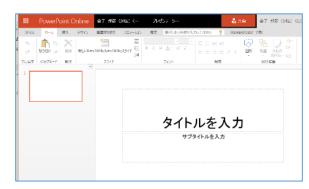
Office 365 の主な機能



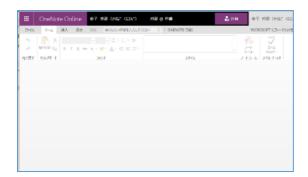




エクセル (表計算)



パワーポイント (プレゼン)



ワンノート(電子ノート) アウトルック(電子メール)

パソコンでレポートを作成したり,発表したり, データをまとめたりで便利

Office 365 の種類

・Office 365 のオンライン版

WEBブラウザで使う.

https://portal.office.com

各自の **ID** と**パスワード**でサインインが必要.

Office 365 のアプリ版

前もってインストールが必要。

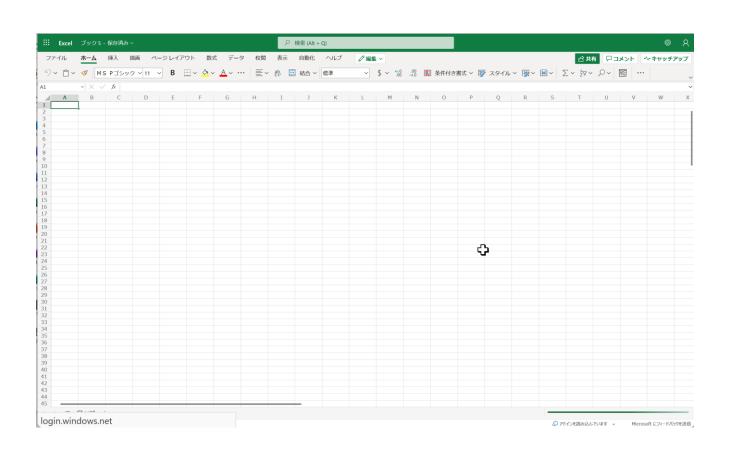
インストールでは、大量の通信が行われる.

(時間がかかる. 通信費用にも注意)

2種類ある. この授業では, **どちらを使用しても問題** ない

【要点】 Web ブラウザで、次のページを開き、各自の ID とパスワードでサインイン

https://portal.office.com



① Web ブラウザで,次のページを開く https://portal.office.com

② **電子メールアドレス**を入れる.「**次へ**」をクリック. **(例)p1234567**@fukuyama-u.ac.jp



③ **パスワード**を入れ,「**サインイン**」を**クリック** パスワードは,各自が設定したもの



サインイン

④ Excel を使いたいときは、メニューで Excel を選ぶ







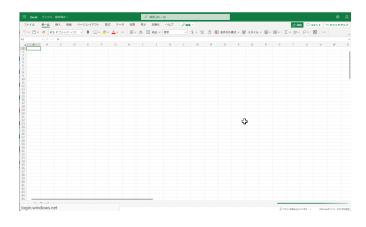
さまざまなメニュー

⑤ Excel のブックの種類を選ぶ



この授業では「新しい空白のブック」を使う

⑥ Excel の画面が開く



まとめ

① 統計データの検索

分野での検索:**企業・家計・経済→家計調査**

キーワード検索:家計調査

② 必用な項目を選ぶ (表示項目選択)

品目分類:プリン

時間軸(月次):最新月

③ レイアウトの設定

ページ, 行, 列の設定

④ ダウンロード

演習

・各自で、e Stat 家計調査データで、「プリン」や、 それ以外の品目について調べてみる

• Excel を使い, ヒストグラムを作成してみる

3-2 クロス集計表

ピボットテーブル(クロス集計表)の例

名前 A B C D E	性別 男性 男性 女性 女性 男性 男性	申し込み 済 未 済 未 済 未	
-----------------------------	--	------------------------------------	--

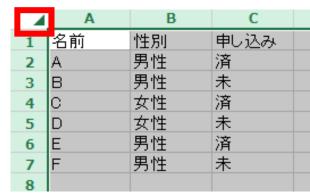
	男性	女性
済	2	1
未	2	1

元データ

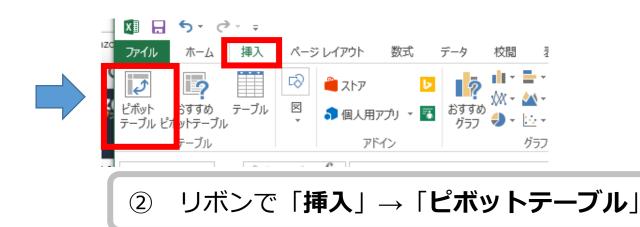
ピボットテーブル (クロス集計表)の例

Excel でピボットテーブル(クロス集計用) を作成(1/5)

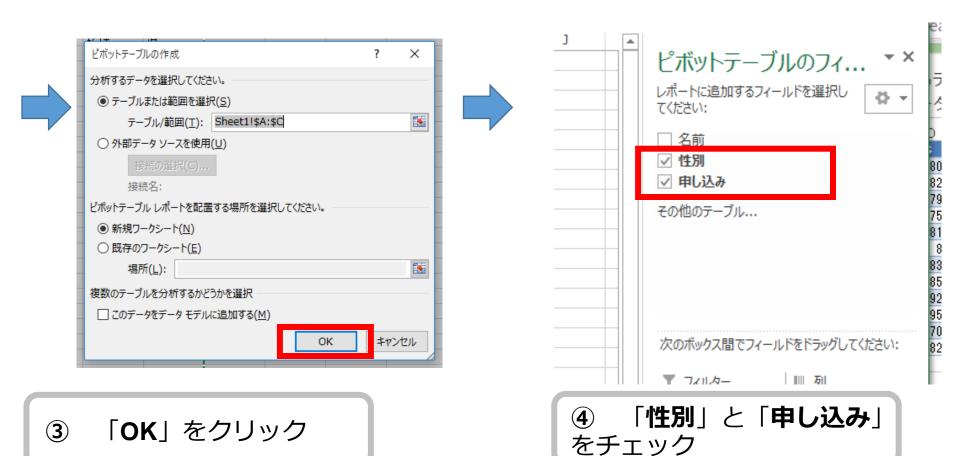




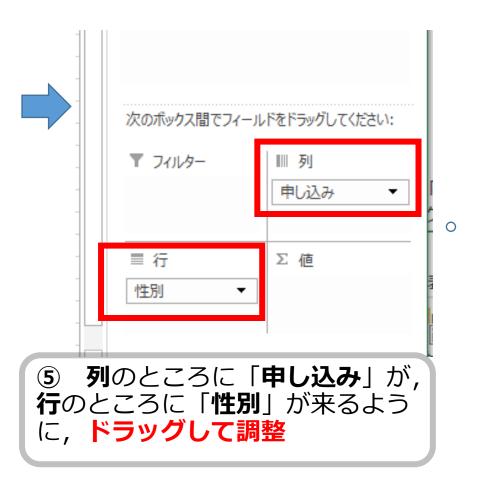
左上を**クリック**.全セルが選択される
 (これは**範囲の選択**)



Excel でピボットテーブル(クロス集計用) を作成(2/5)



Excel でピボットテーブル(クロス集計用) を作成(3/5)

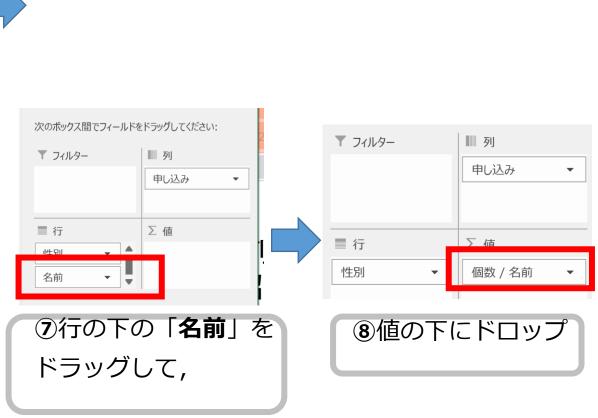




Excel でピボットテーブル(クロス集計用) を作成(4/5)



⑥「名前」をチェック



Excel でピボットテーブル(クロス集計用) を作成(5/5)



ピボットテーブル(クロス集計表)ができた

クロス集計表の例

最終卒業学校卒業のあと,**引っ越しをしていない**のは全体の何パーセントか?

最終卒業学校卒業時の居住地(男女別)

	まだ卒業し ていない	現在と同じ 居住地	その他
男性	19.2	17.8	63.0
女性	17.3	10.6	72.1

パーセント

ここから導かれる結論は? **性別と居住地についての関連** その他: 居住地が違う,不詳

e-Stat 社会保障・人口問題基本調査(人口移動調査) / 人口 移動調査 / 第8回人口移動調査 より

3-3 相関

相関

相関は、2つの変数の間に関連性があるか (一方が変化すれば、もう一方も変化する関係)

・相関あり

X が増えると、Yが増えている

Xが増えると、Yが減っている

・相関なし

XとYに関係がない

相関係数

• 相関係数は, 相関を算出した数値

1や - 1に近い値: 相関あり

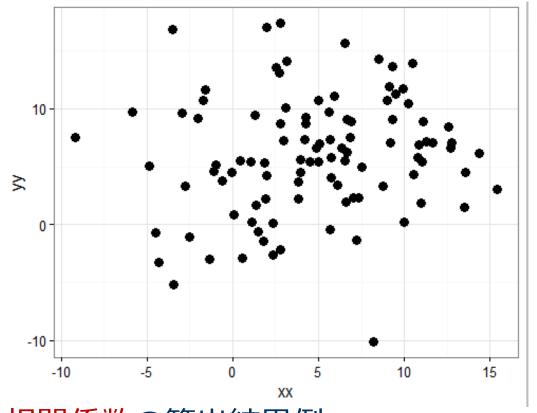
1に近い値: 正の相関関係

- 1 に近い値: **負**の**相関**関係

0 に近い値: 相関なし

2つの変数の例

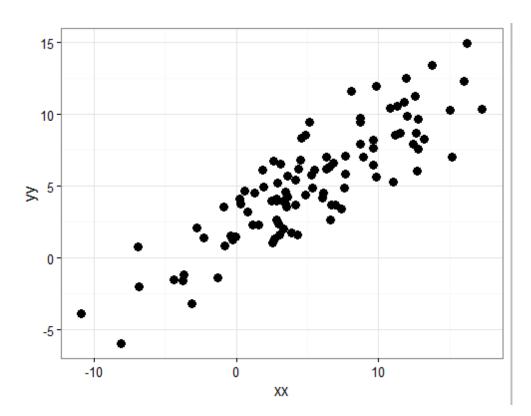
2つの変数 xx, yy の散布図



相関係数の算出結果例

0.1252164 (0 に近い値)

2つの変数の例



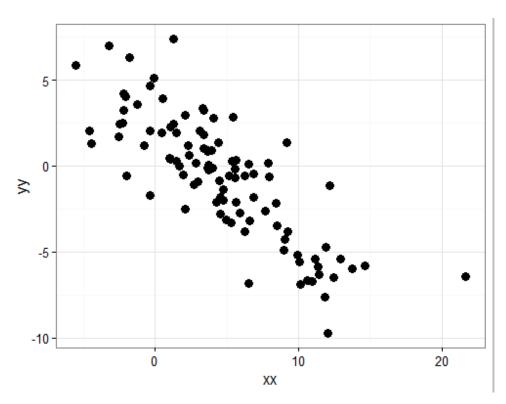
2つの変数 xx, yy が 互いに相関関係を持 つ場合.

xx の値が<u>増える</u>と yy の値が<u>増える</u>という 正の相関関係

相関係数の算出結果例

0.8620027 (1 に近い値)

2つの変数の例



2つの変数 xx, yy が 互いに相関関係を持 つ場合.

xx の値が<u>増える</u>と yy の値が<u>減る</u>という 負の相関関係

相関係数の算出結果例

-0.8502535 (-1 に近い値)

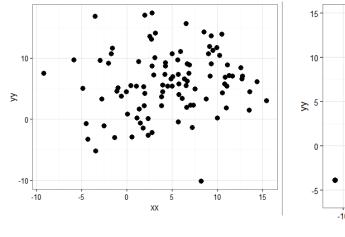
相関係数

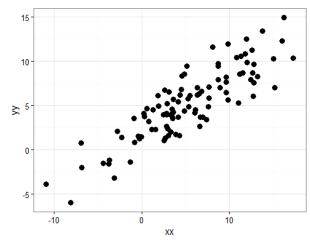
・1や -1に近い値: 相関あり

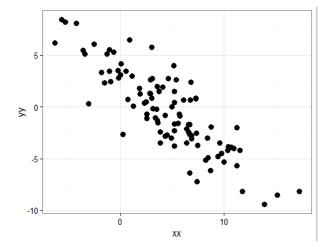
1に近い値: 正の相関関係

- 1に近い値: 負の相関関係

0に近い値: 相関なし







0.1252164 (0 に近い値)

-0.8502535 (-1 に近い値)

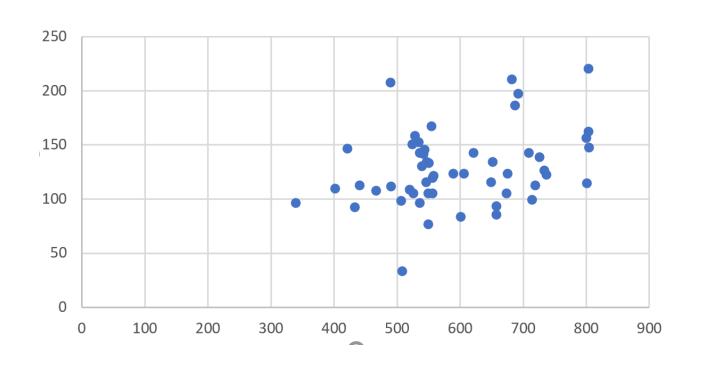
0.8620027 (1 に近い値)

正の相関関係

負の相関関係 43

e Stat 家計調査のデータを用いた分析

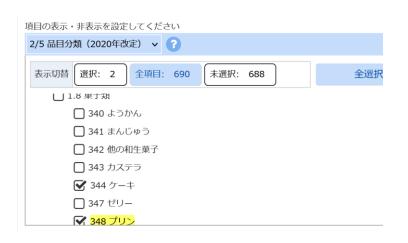
「**プリン**」にお金を使うことと,「**ケーキ**」に お金を使うことに**相関**があるか?



横軸:ケーキ、縦軸:プリン

e Stat 家計調査のデータの準備

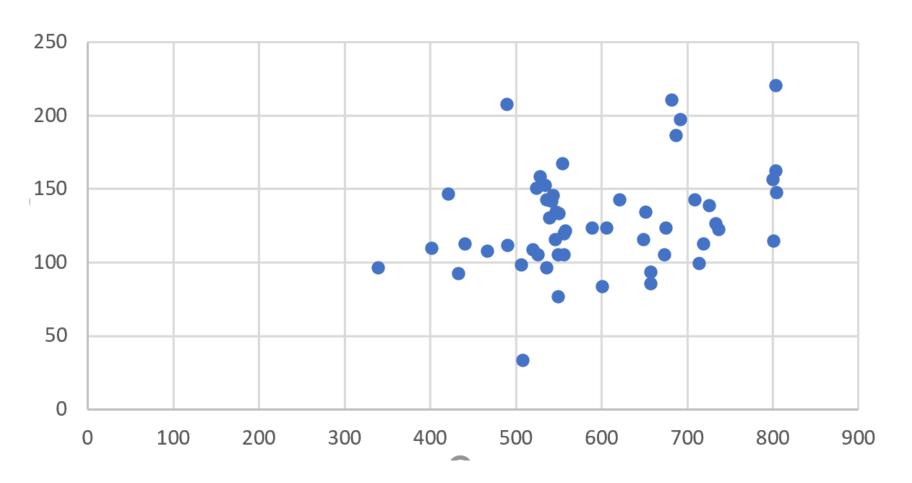
• e Stat で,表示項目として,**ケーキ**とプリンを選択



• データをダウンロード, Excel で開く

	A	В	C	D	E	F	G	н	1	J	r.
2	表章項目	表章項目	世帯区分	世帯区分	時間軸(月	時間軸(月	地域区分	地域区分	/品目分類	344 ケーキ	348 プリン
3	1	金額	3	二人以上の	2E+09	#######	0	全国		551	133
4	1	金額	3	二人以上の	2E+09	#######	1003	01100 札(晃市	529	158
5	1	金額	3	二人以上の	2E+09	#######	2003	02201 青	た 市	340	96
6	1	金額	3	二人以上の	2E+09	#######	3003	03201 盛	岡市	804	162
7	1	金額	3	二人以上の	2E+09	#######	4003	04100 仙	台市	658	85
8	1	金額	3	二人以上の	2E+09	#######	5003	05201 秋日	田市	715	99
9	1	金額	3	二人以上の	2E+09	#######	6003	06201 山	形市	441	112
0	1	金額	3	二人以上の	2E+09	#######	7003	07201 福.	島市	527	105
1	1	金額	3	二人以上の	2E+09	#######	8003	08201 水	市	537	142
2	1	金額	3	二人以上の	2E+09	#######	9003	09201 宇	邪宮市	490	207
3	1	金額	3	二人以上の	2E+09	#######	10003	10201 前標	喬市	607	123
4	1	金額	3	二人以上の	2E+09	#######	11003	11100 さ	ハたま市	683	210
5	1	金額	3	二人以上の	2E+09	#######	12003	12100 千	薬市	805	147
6	1	金額	3	二人以上の	2E+09	#######	13003	13100 東	京都区部	550	105
7	1	金額	3	二人以上の	2E+09	#######	14003	14100 横	兵市	535	152
8	1	金額	3	二人以上の	2E+09	#######	15003	15100 新	烏市	540	130
9	1	金額	3	二人以上の	2E+09	#######	16003	16201 富	山市	544	145
0	1	金額	3	二人以上の	2E+09	#######	17003	17201 金	尺市	804	220
1	1	金額	3	二人以上の	2E+09	#######	18003	18201 福	井市	557	119
2	1	金額	3	二人以上の	2E+09	#######	19003	19201 甲/	存市	650	115
3	1	金額	3	二人以上の	2E+09	#######	20003	20201 長	野市	738	122

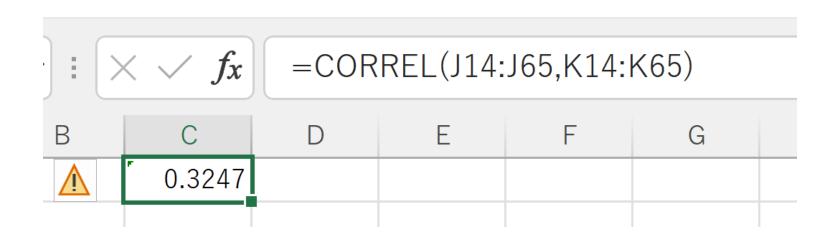
グラフ タイトル



横軸:ケーキ、縦軸:プリン

Excel で相関係数を求める

- ① 相関係数を求めたいセルを選んでおく
- ② メニューの「数式」
- ③「その他の関数」, 「統計」, 「CORREL」
- ④ 配列1と配列2の範囲を指定



相関係数として 0.3247 が求まった → 「相関がある」とは言い切れない

3-4 誤差, 平均



①同じものを計測する. 但し、誤差を含む

②繰り返し計測し、平均をとる

計測値 = 全体の平均 + 誤差 という考え方も

③多数の計測の繰り返しが重要

3-5 オープンデータ

ビックデータによる課題解決

- ・膨大な量のデータ(ビッグデータ)を活用. 課題 解決,効率化,透明化の向上を達成
- プライバシ保護は当然
- 公益性の高いデータはオープンデータとして公開 される

新サービスの例



不動産販売価格予測サイト GEEO https://geeo.otani.co

オープンデータ

① 営利目的, 非営利目的を問わず二次利用可能なルールがされたもの

②機械判読に適したもの

③ 無償で利用できるもの

・総務省のオープンデータの基本指針より