

Working with statistics

Tipps, Dos & Don'ts

Takeru Shibayama

I'm going to talk about...

Use of statistical method – with “real” examples

- Descriptive statistics with MS Excel
- Indexation: adjusting values
- Other types of adjustments
- Stratification – dividing data into right groups
- Two-dimensional data – scatter plot
 - Correlation & Causality

Descriptive statistics with MS Excel



- MS Excel: a powerful tool with many built-in statistical functions.
 - You don't have to do any complex calculation by hand!
 - Spreadsheet can handle a large number of data!
 - Chart with a few clicks – easy to visualize the data.
- Explanation in the next slides: based on Excel 2013 or newer
 - Including Office 365

I'm going to talk about...

Use of statistical method – with “real” examples

- Descriptive statistics with MS Excel
- Indexation: adjusting values
- Other types of adjustments
- Stratification – dividing data into right groups
- Two-dimensional data – scatter plot
 - Correlation & Causality

A sample data (1)

- ... some dataset you've never seen?
- Price of soy source in Japan
- The consumer price survey
 - Data available from <https://www.e-stat.go.jp/>
 - Released on 23 October 2020



5 November 2020



INTERACT Online Event

	A	B
1	Place	Price October 2020
2	Sapporo	¥ 311
3	Hakodate	¥ 285
4	Asahikawa	¥ 277
5	Aomori	¥ 271
6	Hachinohe	¥ 271
7	Morioka	¥ 275
8	Sendai	¥ 246
9	Akita	¥ 333
10	Yamagata	¥ 289
11	Fukushima	¥ 242
12	Koriyama	¥ 325
13	Mito	¥ 228
14	Hitachi	¥ 235
15	Utsunomiya	¥ 273
16	Oyama	¥ 246
17	Maebashi	¥ 236
18	Maebashi	¥ 236

Source: <https://tamariva.com/> <https://sakurashoyu.com/>

A sample data (2)

- Data from 81 places
- What is:
 - Maximum?
 - Minimum?
 - Mean? (Average?)
 - Median?
 - Standard deviation?

1	Place	Price October 2020	38	A	B	74	Kitakyushu	¥	314
2	Sapporo	¥ 311	39	Kofu	¥ 225	75	Saga	¥	273
3	Hakodate	¥ 285	40	Nagano	¥ 281	76	Nagasaki	¥	324
4	Asahikawa	¥ 277	41	Matsumoto	¥ 236	77	Sasebo	¥	264
5	Aomori	¥ 271	42	Gifu	¥ 246	78	Kumamoto	¥	253
6	Hachinohe	¥ 271	43	Shizuoka	¥ 249	79	Oita	¥	296
7	Morioka	¥ 275	44	Hamamatsu	¥ 242	80	Miyazaki	¥	280
8	Sendai	¥ 246	45	Fuji	¥ 232	81	Kagoshima	¥	328
9	Akita	¥ 333	46	Nagoya	¥ 273	82	Naha	¥	293
10	Yamagata	¥ 289	47	Okazaki	¥ 253				
11	Fukushima	¥ 242	48	Tsu	¥ 278				
12	Koriyama	¥ 325	49	Matsusaka	¥ 253				
13	Mito	¥ 228	50	Otsu	¥ 228				
14	Hitachi	¥ 235	51	Kyoto	¥ 269				
15	Utsunomiya	¥ 273	52	Osaka	¥ 241				
16	Oyama	¥ 246	53	Sakai	¥ 221				
17	Maebashi	¥ 236	54	Hirakata	¥ 246				
18	Saitama	¥ 221	55	Higashiosaka	¥ 231				
19	Kumagaya	¥ 185	56	Kobe	¥ 214				
20	Kawaguchi	¥ 217	57	Himeji	¥ 260				
21	Tokorozawa	¥ 253	58	Nishinomiya	¥ 233				
22	Chiba	¥ 235	59	Itami	¥ 217				
23	Sakura	¥ 231	60	Nara	¥ 236				
24	Urayasu	¥ 253	61	Wakayama	¥ 282				
25	Ku-area of Tokyo	¥ 246	62	Tottori	¥ 301				
26	Hachioji	¥ 215	63	Matsue	¥ 285				
27	Tachikawa	¥ 203	64	Okayama	¥ 260				
28	Fuchu	¥ 309	65	Hiroshima	¥ 302				
29	Yokohama	¥ 230	66	Fukuyama	¥ 253				
30	Kawasaki	¥ 230	67	Yamaguchi	¥ 308				
31	Sagamihara	¥ 250	68	Ube	¥ 289				
32	Yokosuka	¥ 234	69	Tokushima	¥ 311				
33	Niigata	¥ 253	70	Takamatsu	¥ 260				
34	Nagaoka	¥ 264	71	Matsuyama	¥ 248				
35	Toyama	¥ 254	72	Imabari	¥ 264				
36	Kanazawa	¥ 235	73	Kochi	¥ 231				
37	Fukui	¥ 286		Fukuoka	¥ 260				
38	Kofu	¥ 225							

Excel Functions for Descriptive Statistics

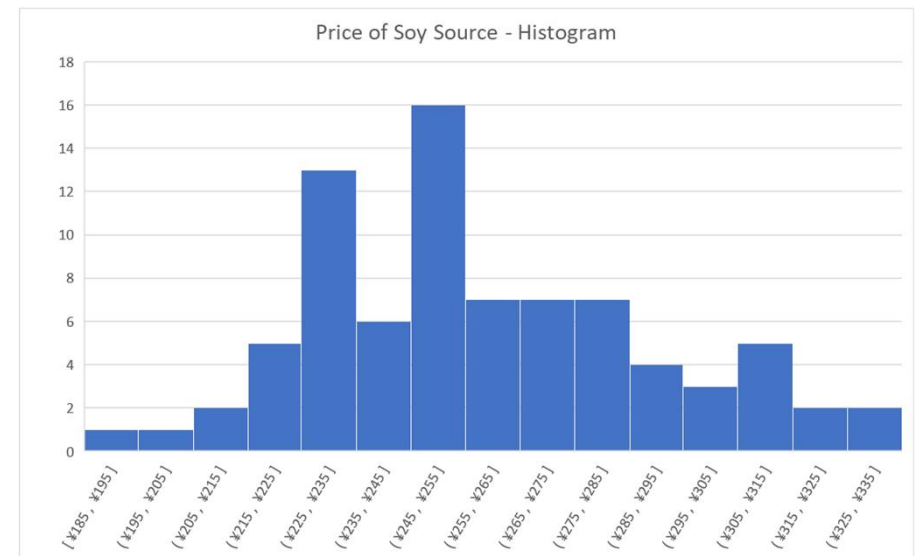
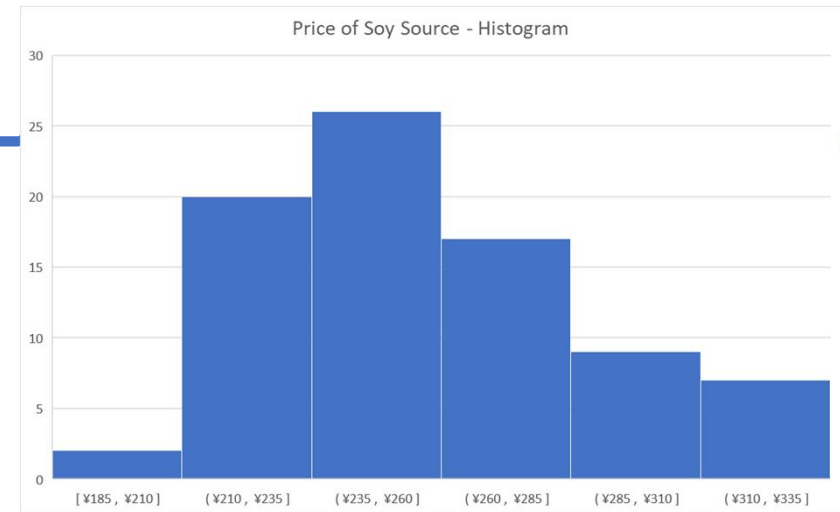
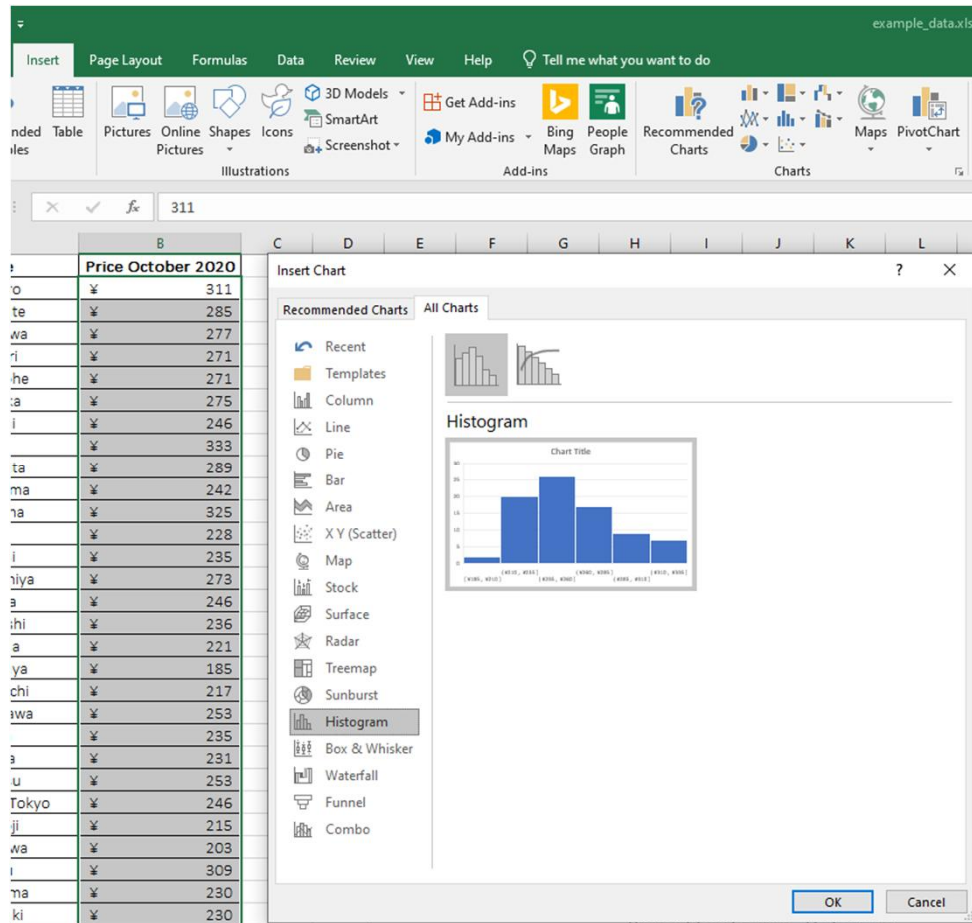
- Maximum: MAX
- Minimum: MIN
- Mean (Average): AVERAGE
- Median: MEDIAN
- Standard deviation: STDEV.P and STDEV.S
 - Your data = Entire population → use STDEV.P
 - Your data = Sample of your entire population → STDEV.S
 - In many case: your data is a sample

	A	B	C
83			
84	Max.	333	
85	Min.	185	
86	Mean Value (AVERAGE)	=AVERAGE(B2:B82)	
87	Median	253	
88	Std. Deviation	31.5	

$$\sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

$$\sqrt{\frac{\sum (x - \bar{x})^2}{(n-1)}}$$

Histogram



I'm going to talk about...

Use of statistical method – with “real” examples

- Descriptive statistics with MS Excel
- Indexation: adjusting values
- Other types of adjustments
- Stratification – dividing data into right groups
- Two-dimensional data – scatter plot
 - Correlation & Causality

Another sample data

- Include both
 - October 2020 data and
 - October 2019 data
- What is the problem with these two datasets?
- The value of 1 yen this year and last year is different!
- Adjustment by consumer price index

1	Place	Price October 2020	Price October 2019
2	Sapporo	¥ 311	¥ 305
3	Hakodate	¥ 285	¥ 301
4	Asahikawa	¥ 277	¥ 278
5	Aomori	¥ 271	¥ 271
6	Hachinohe	¥ 271	¥ 275
7	Morioka	¥ 275	¥ 246
8	Sendai	¥ 246	¥ 294
9	Akita	¥ 333	¥ 333
10	Yamagata	¥ 289	¥ 289
11	Fukushima	¥ 242	¥ 242
12	Koriyama	¥ 325	¥ 302
13	Mito	¥ 228	¥ 269
14	Hitachi	¥ 235	¥ 235
15	Utsunomiya	¥ 273	¥ 273
16	Oyama	¥ 246	¥ 241
17	Maebashi	¥ 236	¥ 236
18	Saitama	¥ 221	¥ 250
19	Kumagaya	¥ 185	¥ 213
20	Kawaguchi	¥ 217	¥ 213
21	Tokorozawa	¥ 253	¥ 235
22	Chiba	¥ 235	¥ 235
23	Sakura	¥ 231	¥ 239
24	Urayasu	¥ 253	¥ 269
25	Ku-area of	¥ 246	¥ 249
26	Hachioji	¥ 215	¥ 220
27	Tachikawa	¥ 203	¥ 199
28	Fuchu	¥ 309	¥ 299
29	Yokohama	¥ 230	¥ 237
30	Kawasaki	¥ 230	¥ 227
31	Sagamihara	¥ 250	¥ 246

Adjustment with index – indexation (1)

- Consumer price index Latest available data: September 2020
 - October price index is not yet available!
 - Alternatively, let's use September indexes from 2020 and 2019!
- Consumer price indexes:
 - (in case of Japan: 2015 price = 100)
 - September 2019: 105.2
 - September 2020: 107.2
- Formula for adjustment:

$$P_{adj} = P_{orig} \times \frac{Index_{20}}{Index_{19}}$$

Adjustment with index – indexation (2)

- Hint: prepare a table with the indexes
- “Translate” the formula into the Excel style

	G	H	I
	Month	Index	
	Sep-19	105.2	
	Sep-20	107.2	

$$P_{adj} = P_{orig} \times \frac{Index_{20}}{Index_{19}}$$

↓
↓

Column D Column C

Fixed: H3
 Fixed: H2

$$=C2 * \$H\$3 / \$H\$2$$

Type into Row 2

	A	B	C	D	E	F	G	H
1	Place	Price October 2020	Price October 2019	Price October 2019 after Indexation			Month	Index
2	Sapporo	¥ 311	¥ 305	=C2*(\$H\$3/\$H\$2)			Sep-19	105.2
3	Hakodate	¥ 285	¥ 301	¥ 307			Sep-20	107.2

Copy into Row 3, 4, ...

	A	B	C	D	E	F	G	H
1	Place	Price October 2020	Price October 2019	Price October 2019 after Indexation			Month	Index
2	Sapporo	¥ 311	¥ 305	¥ 311			Sep-19	105.2
3	Hakodate	¥ 285	¥ 301	=C3*(\$H\$3/\$H\$2)			Sep-20	107.2

Tipps: Excel Operator “\$”

- Without Operator “\$”: relative reference
 - Formula is adjusted when copied into another cell
- With Operator “\$”: absolute reference
 - Formula is NOT adjusted when copied into another cell
- Applicable for both columns and rows
 - A1 → Both columns and rows are adjusted.
 - A\$1 → Only rows are fixed & columns are adjusted.
 - \$A1 → Only columns are fixed & rows are adjusted.
 - \$A\$1 → Both columns and rows are not adjusted.
- Previous example: reference to fixed cells for 2019 and 2020 price index → Dollar operator for both columns and rows (\$A\$1)

Adjustment with index – indexation (2)

- Descriptive statistics per dataset:

Descriptive Statistics	Price October 2020	Price October 2019	Price October 2019 after Indexation
Max.	333	355	362
Min.	185	199	203
Mean (AVERAGE)	259	265	270
Median	253	264	269
Std. Deviation	32	32	33

- Oct 2020 and Oct 2019 price WITHOUT indexation

Max.	355
Min.	185
Mean	262
Median	260
Std. Deviation	32

- Oct 2020 and Oct 2019 price WITH indexation

Max.	362
Min.	185
Mean	265
Median	260
Std. Deviation	32

Other possible adjustment/indexation

- Wage
- Social security contribution by employer
- Tax
- Exchange rate
- ...

I'm going to talk about...

Use of statistical method – with “real” examples

- Descriptive statistics with MS Excel
- Indexation: adjusting values
- Other types of adjustments
- Stratification – dividing data into right groups
- Two-dimensional data – scatter plot
 - Correlation & Causality

Data Stratification (1)

- Stratification (stratified sampling):
 - Sampling data from each subpopulation (*stratum*)
 - When each subpopulation has:
 - Different characteristics, and/or
 - Very different distribution (or descriptive statistics)
 - *Strata*: definition of division of population
 - Collectively exhaustive: covering all *stratum* altogether
 - Mutually exclusive: no sample in two or more *stratum*

Hourly unit cost	Position
€ 62.38	Manager
€ 29.18	Administrative staff
€ 52.34	Engineer
€ 24.69	Administrative staff
€ 78.11	Manager
€ 32.33	Student Employee
€ 19.82	Administrative staff
€ 42.25	Administrative staff
€ 34.10	Engineer
€ 39.38	Administrative staff
€ 32.31	Engineer
€ 40.01	Engineer
€ 26.31	Administrative staff
€ 19.01	Student Employee
€ 35.21	Administrative staff
€ 15.21	Engineer
€ 14.39	Student Employee
€ 13.99	Student Employee
€ 35.19	Administrative staff
€ 51.04	Engineer
€ 90.32	Consultant
€ 48.42	Consultant
€ 46.21	Engineer
€ 27.89	Administrative staff
€ 54.32	Manager
€ 62.18	Engineer
€ 32.33	Administrative staff
€ 39.61	Engineer
€ 28.39	Administrative staff
€ 33.01	Administrative staff

Data Stratification (2)

- Good examples of “strata”:
 - Male and Female
 - Full-time and part-time employees at Company X
 - Children, Adults and Elderly people
 - People in Country A and People in Country B, if you’re dealing only with these two countries
- Bad examples of “strata”
 - People in Country A and People in Country B, even if you have to deal with Countries C and D, too. (Collectively NOT exhaustive!)
 - Employee in maternity leave, School employee, Civil Servants (Mutually NOT exclusive!)

Data Stratification (3) – Sample size allocation

- Example: Your historical data over 7 years has an hourly salary of 3240 persons in the dataset, and it is known that:
 - 423 persons are managers
 - 1924 persons are administrative staffs
 - 742 persons are engineers or consultants
 - 151 persons are student assistants or trainees
- Percentages of these 4 groups are...
 - % manager: $423/3240 = 13.06\%$
 - % administrative staffs: $1924/3240 = 59.38\%$
 - % engineers and consultants: $742/3240 = 22.90\%$
 - % student assistants and trainees: $151/3240 = 4.46\%$
- You randomly select your 400 sample from the dataset of the last 3 years ($n = 1239$), but stratified by the following groups:
 - Manager: $400 \times 13.06\% = 52$ people
 - Administrative staffs: $400 \times 59.38\% = 238$ people
 - Engineers and consultants: $400 \times 22.90\% = 92$ people
 - Student assistants and trainees: $400 \times 4.46\% = 18$ people

I'm going to talk about...

Use of statistical method – with “real” examples

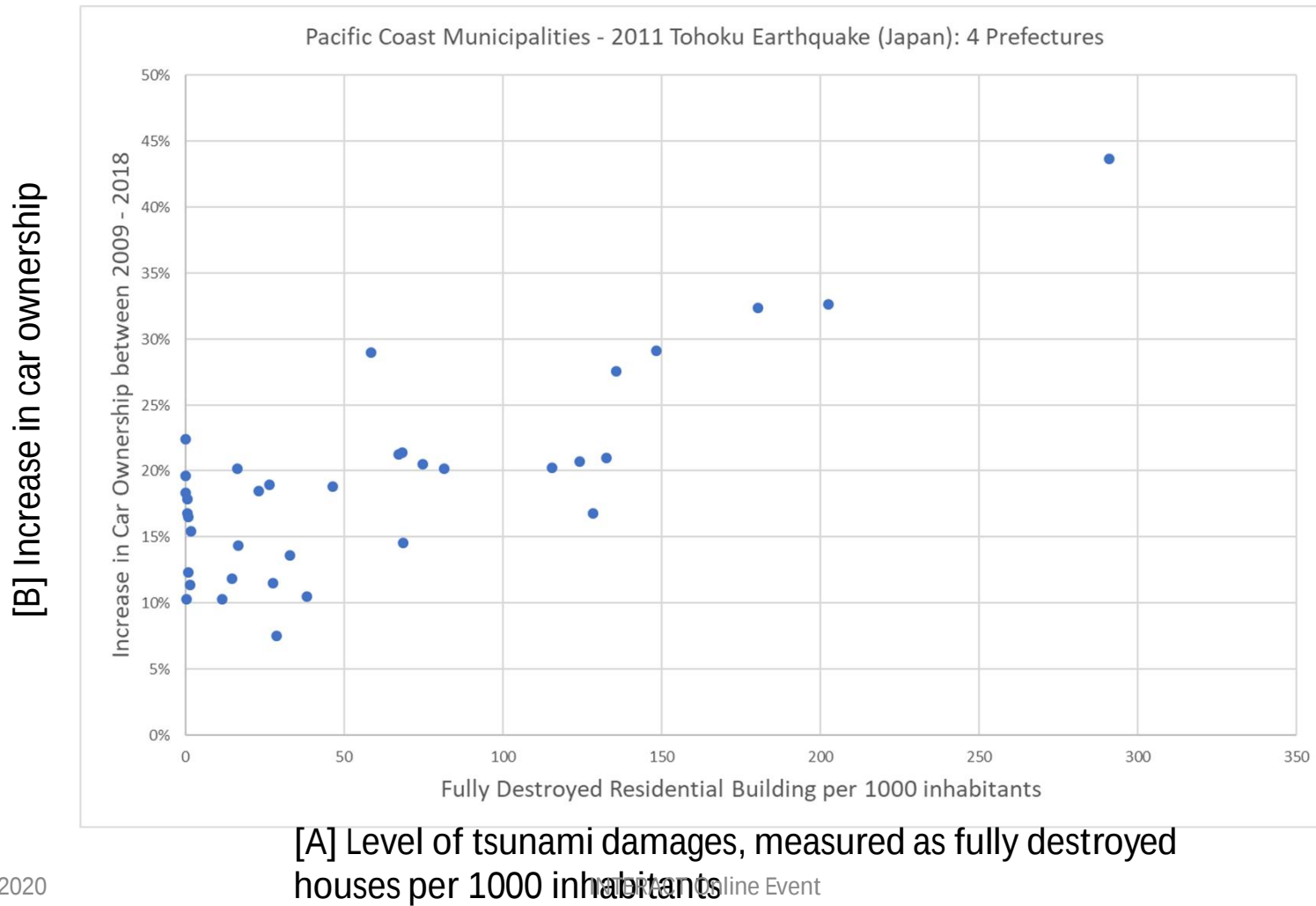
- Descriptive statistics with MS Excel
- Indexation: adjusting values
- Other types of adjustments
- Stratification – dividing data into right groups
- Two-dimensional data – scatter plot
 - Correlation & Causality

Two-dimensional data (1)

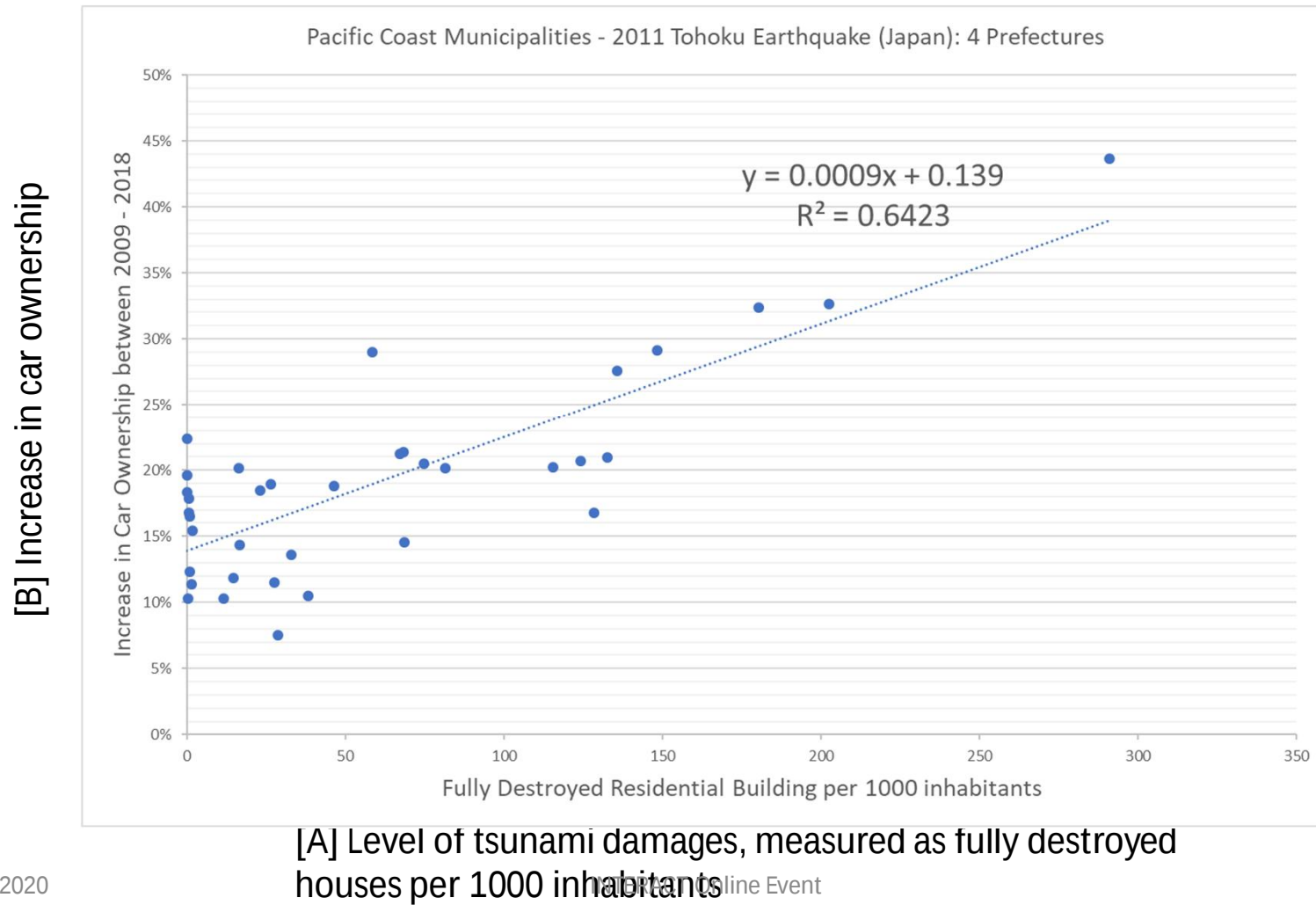
- Two different data around one object, or of two objects
- Examples:
 - [A] Height and [B] weight
 - [A] Car weight and [B] amount of exhaust
 - [A] Temperature and [B] sales of ice cream
- An example from my recent research:
 - [A] Level of tsunami damages, measured as fully destroyed houses per 1000 inhabitants
 - [B] Increase in car ownership



Two-dimensional data (2)



Two-dimensional data (3): Linear Regression



Two-dimensional data (4)

- Correlation: statistical relationship between two random variables
- Correlation coefficient (*Pearson correlation coefficient*):
 - A statistic measuring linear correlation between two variables
 - Between -1 and 0: Negative correlation
 - 0: no correlation
 - Between 0 and 1: Positive correlation
 - Excel function: CORREL
- Coefficient of determination (R-Squared)
 - Easy way: simply the square of correlation coefficient

Correlation and Causality

- Be careful! Correlation $\neq$ Causality
- Example:
 - When sales of ice cream is very good, there are more murder cases. (Therefore we should prohibit ice cream sales to prevent murders.)
 - What's wrong with this sentence?
 - There's another common cause of these two events: temperature
- Correlation may change over time, too!
 - For example: even if you find a correlation between personnel cost and travel cost based on the 2017-2019 data...
 - This may no longer be valid in 2021 because people will have learned online events through COVID-19 pandemic.

End

Contact: takeru.shibayama@tuwien.ac.at