National Statistics Center

NSTAC

A Case Study of Output Checking in Japan

For Reliable Statistics, with Competent Technology

Yutaka Abe

National Statistics Center

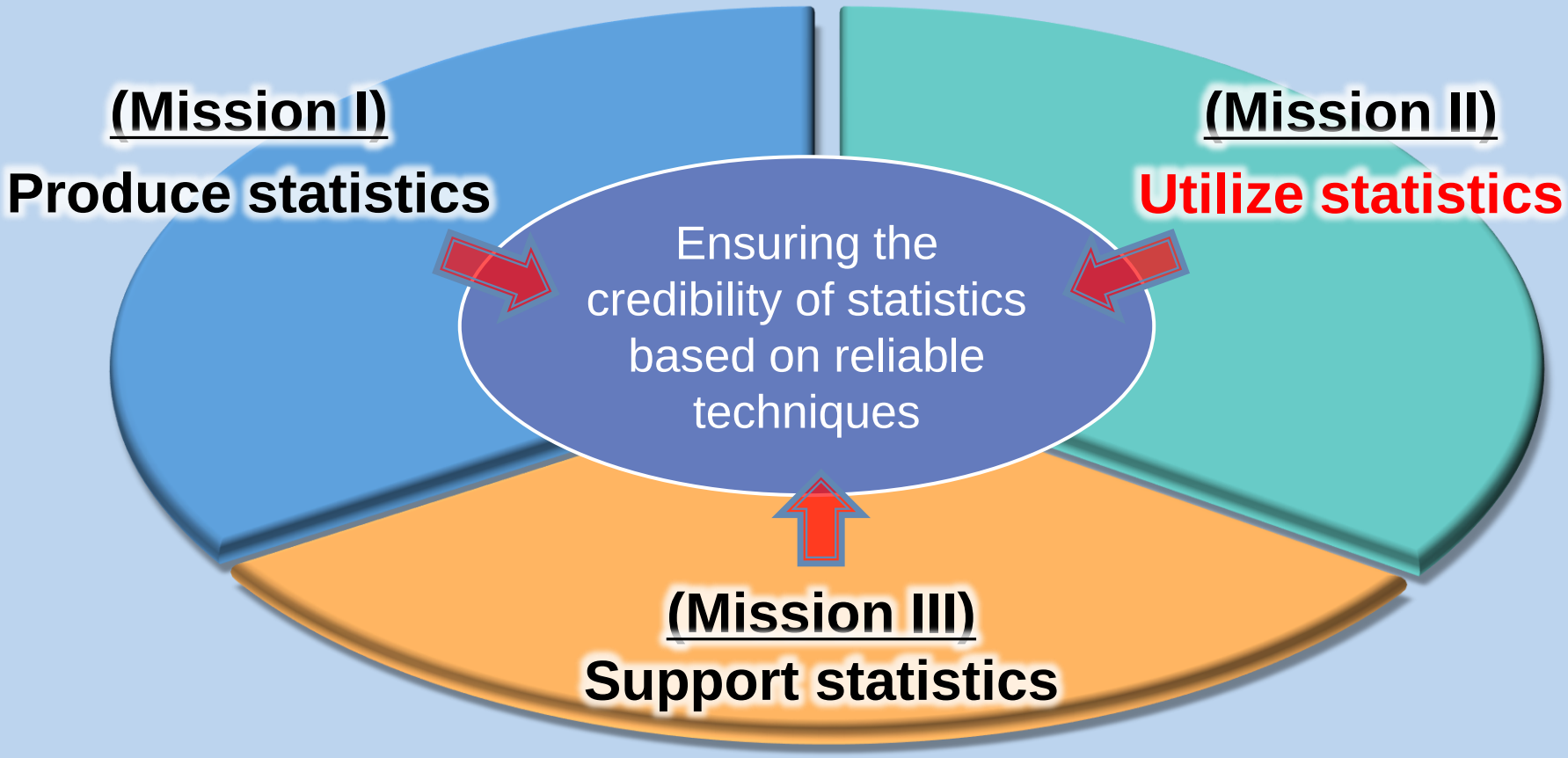
Kazuhiro Minami

The Institute of Statistical Mathematics, National Statistics Center

UNECE Expert Meeting on Statistical Data Confidentiality 2023

2023/9/27

We, NSTAC have 3 Missions



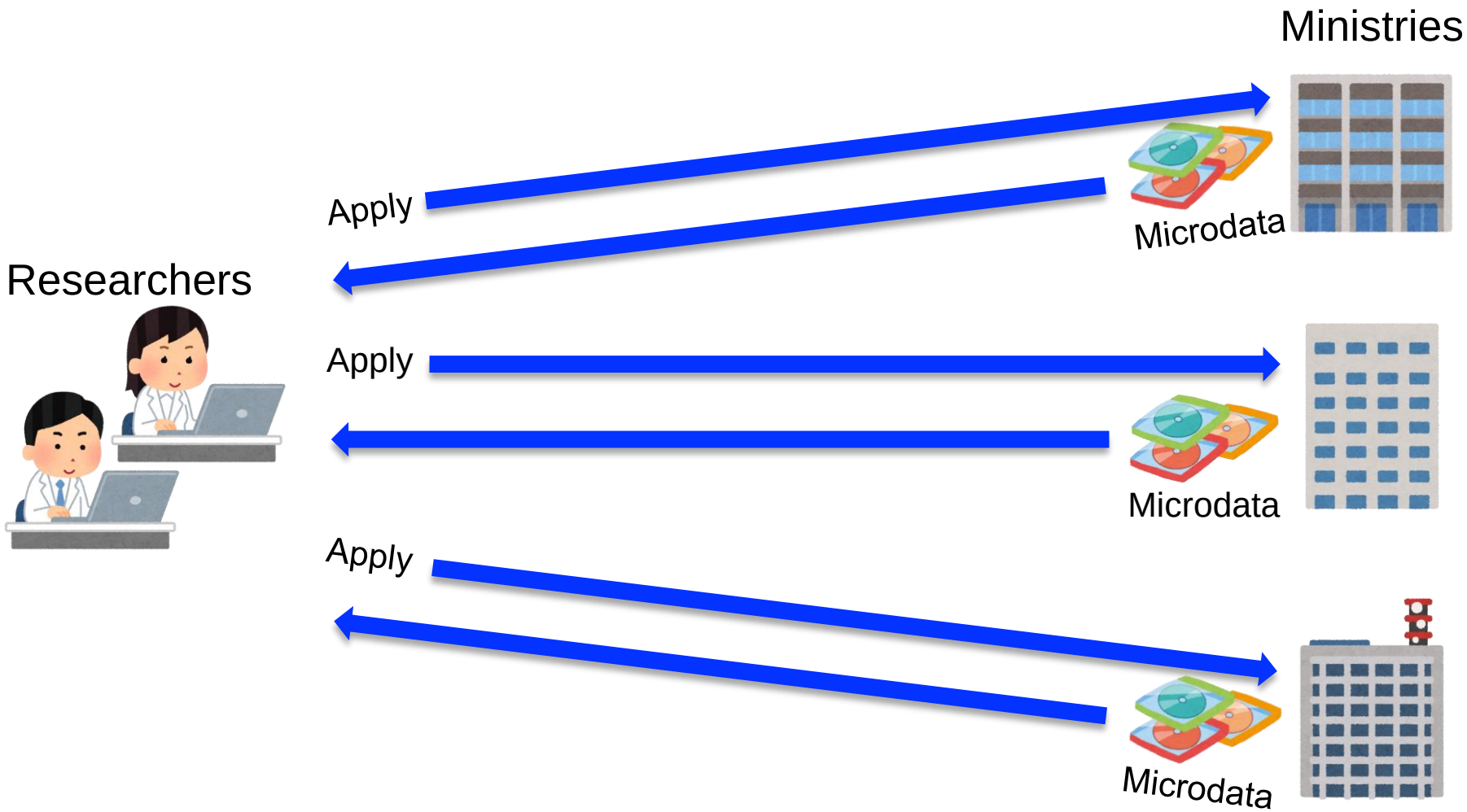
Statistical System of Japan

- Japan has decentralized statistical system
  (multiple government ministries have their own statistical survey)

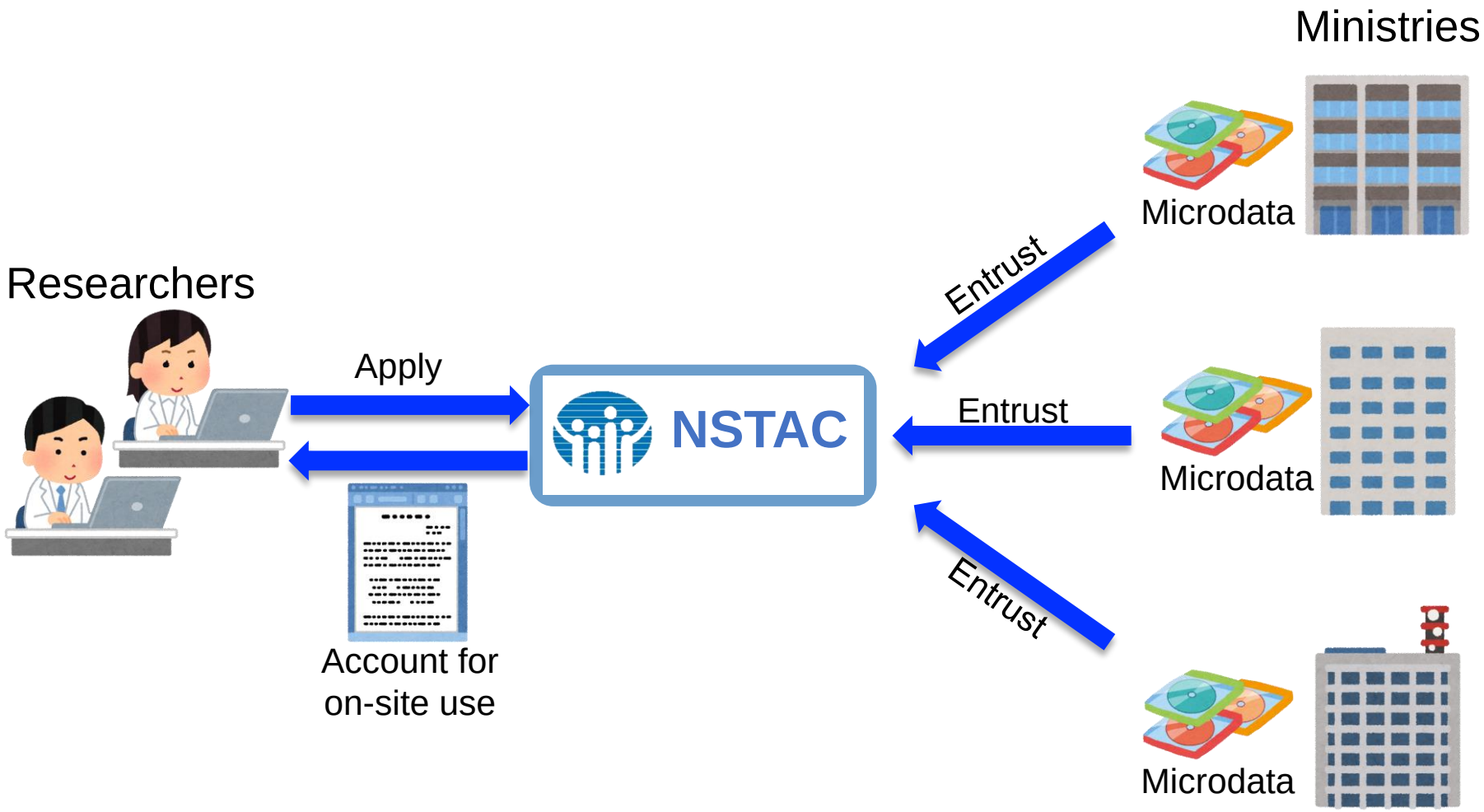
List of the ministries providing the microdata via on-site use

Ministries	Survey Titles
Cabinet Secretariat	Basic Survey on Human Connection
Cabinet Office	Annual Survey of Corporate Behavior, etc.
Children and Families Agency	Survey on the Living of Children, etc.
Ministry of Internal Affairs and Communications	Population Census, Economic Census, Labour Force Survey, Survey on Time Use and Leisure Activities, etc. , etc.
Ministry of Finance	Financial Statements Statistics of Corporations by Industry
Ministry of Education, Culture, Sports, Science and Technology	School Basic Survey, School Teachers Survey, etc.
Ministry of Health, Labour and Welfare	Vital Statistics, Basic Survey on Wage Structure, National Health and Nutrition Survey, etc.
Ministry of Agriculture Forestry and Fisheries	Census of Fisheries, Statistics on Marine Fishery Production
Ministry of Economy, Trade and Industry	Basic Survey of Japanese Business Structure and Activities, Census of Manufacture, etc. , etc.
Ministry of Land, Infrastructure, Transport and Tourism	Statistics on Building Construction Started, Consumption Trend Survey for Foreigners Visiting Japan, etc.
Minister of the Environment	Survey of industrial waste generation and treatment, etc.

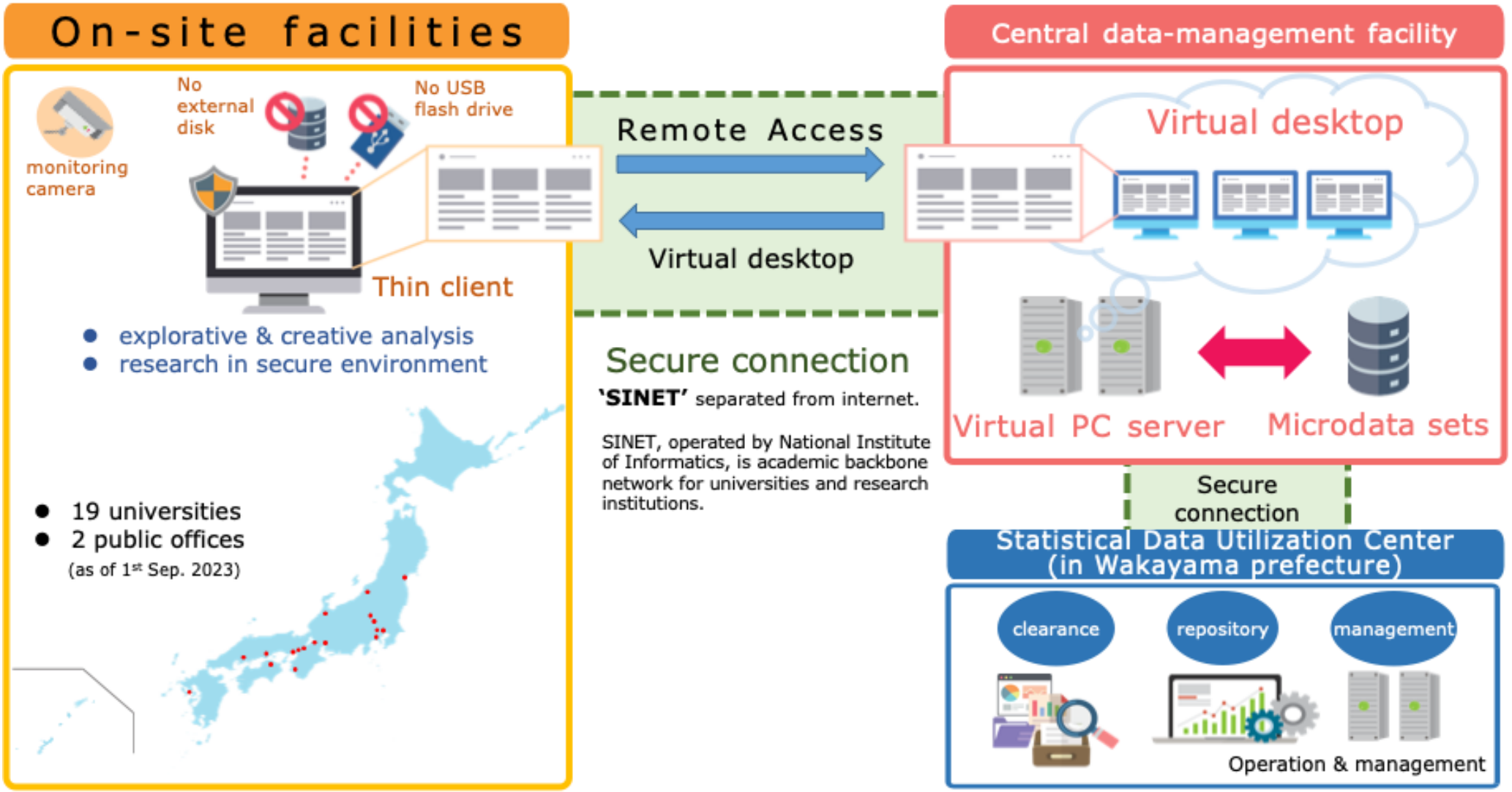
Application of the Microdata via DVD



Application of the Microdata via On-site Use (Since 2019)



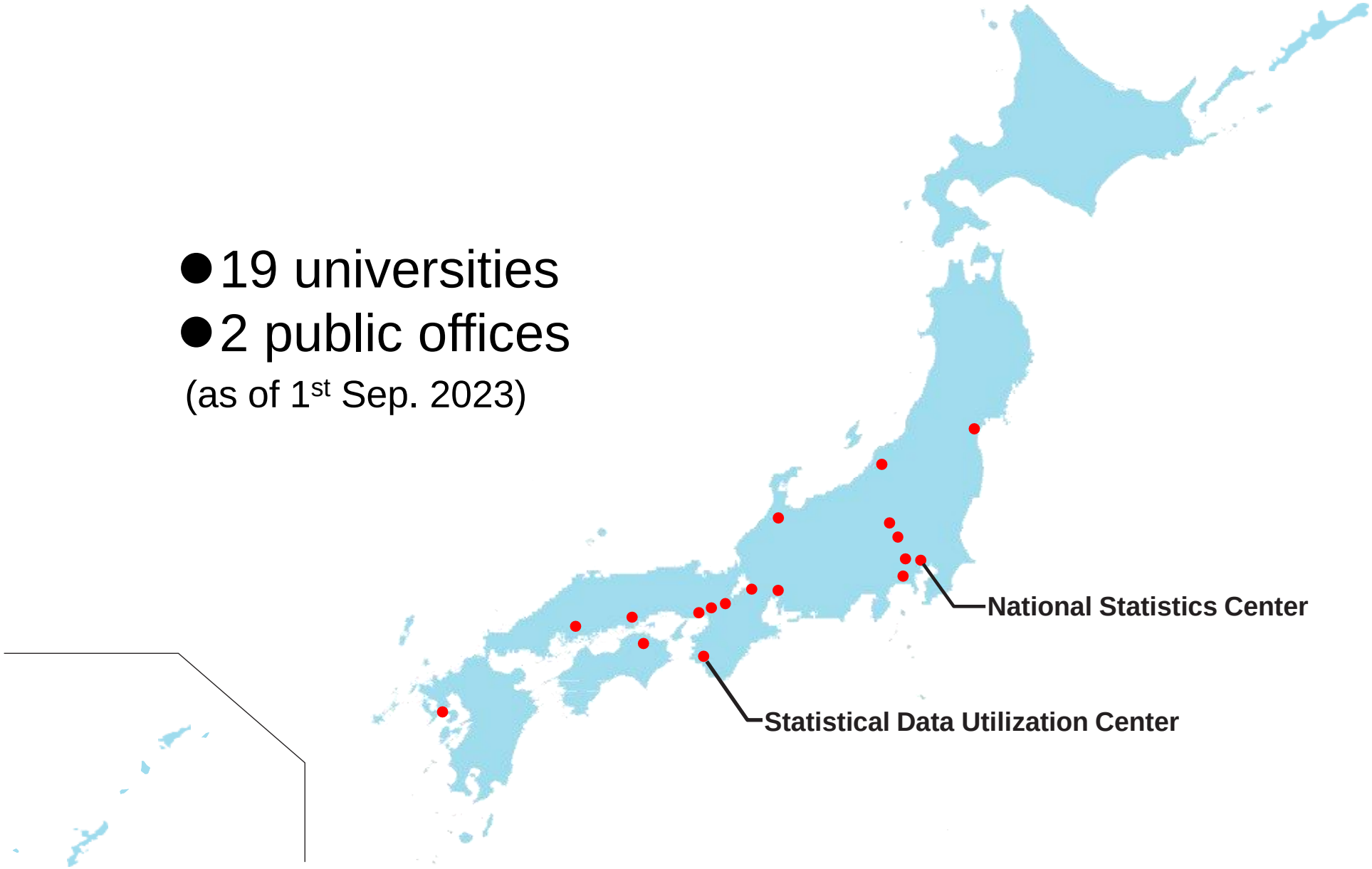
Overview of On-Site System in Japan



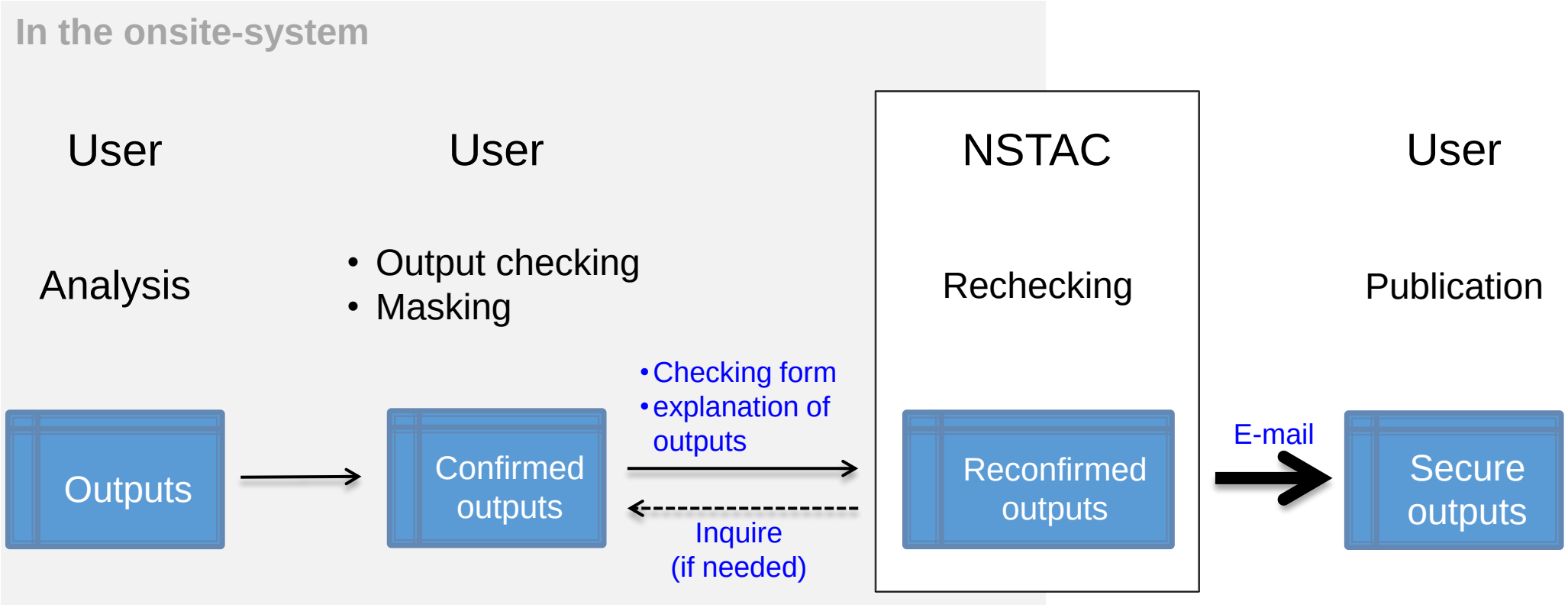
Map of On-Site Facilities in Japan

- 19 universities
- 2 public offices

(as of 1st Sep. 2023)



Output Checking in Japan (1/2)



The output checking rules are defined for commonly used output formats and published as part of the onsite-use manual [1].

Output Checking in Japan (2/2)

- If output format are not included in the output checking rules, we first evaluate if we can handle them with the guiding principles. (Case studies of checking based on the principles are described in the paper.)
- If there are no checking rules for the output, we need to discuss with survey-own ministry about checking methods, so it requires extra time to provide the output.
- For new output formats we frequently encounter, we need to revise the manual to add new rules, to avoid an inquiry to the survey-own ministry.

Consideration of Output Checking Rules for Quartiles

- The median and quartiles are widely used in descriptive statistics, etc.
- Not satisfy the principle of 10 units; their values are calculated from 1 or 2 individuals.
- We have considered Japanese output checking rule before on-site use was launched [2], but we could not find the explicit rules for median and quartiles.
- Generally difficult to accurately infer the rankings of all survey individuals
- It would be possible to establish median and quartile rules by setting proper assumptions.

Case Studies of Data without Boundaries (DwB) Project (1/5)

Project N°: 262608

DwB
Data without Boundaries

ACRONYM: Data without Boundaries

STANDALONE DOCUMENT
Guidelines for Output Checking

WORK PACKAGE 11
(Improved Methodologies for Managing Risks of Access to Detailed OS Data)

Combination of CP & CSA project funded by the European Community
Under the programme "FP7 - SP4 Capacities"
Priority 1.1.3: European Social Science Data Archives and remote access to Official Statistics

Eurostat (2014, August). Guidelines for the checking of output based on microdata research [3].

P.17

T1. If the rank ordering of firms is known or guessable, the percentile cannot be released.

T2. If the variance around the percentile is low, there is the possibility of group disclosure.

T3. If the variance around the percentile is very large, the identity of the percentile respondent might be guessable.

Case Studies of DwB Project (3/5)

T1. If the rank ordering of firms is known or guessable, the percentile cannot be released.

Our Assumption:

It is possible that the top rankings are known or can be inferred, however in general, if the data size is large enough, it is assumed that it is difficult to accurately determine the rankings of individuals near the median and quartiles.

Case Studies of DwB Project (4/5)

T2. If the variance around the percentile is low, there is the possibility of group disclosure.

→ We should introduce an additional rule to prevent group disclosure that would apply to sensitive variables as we do for *sum* and *mean*.

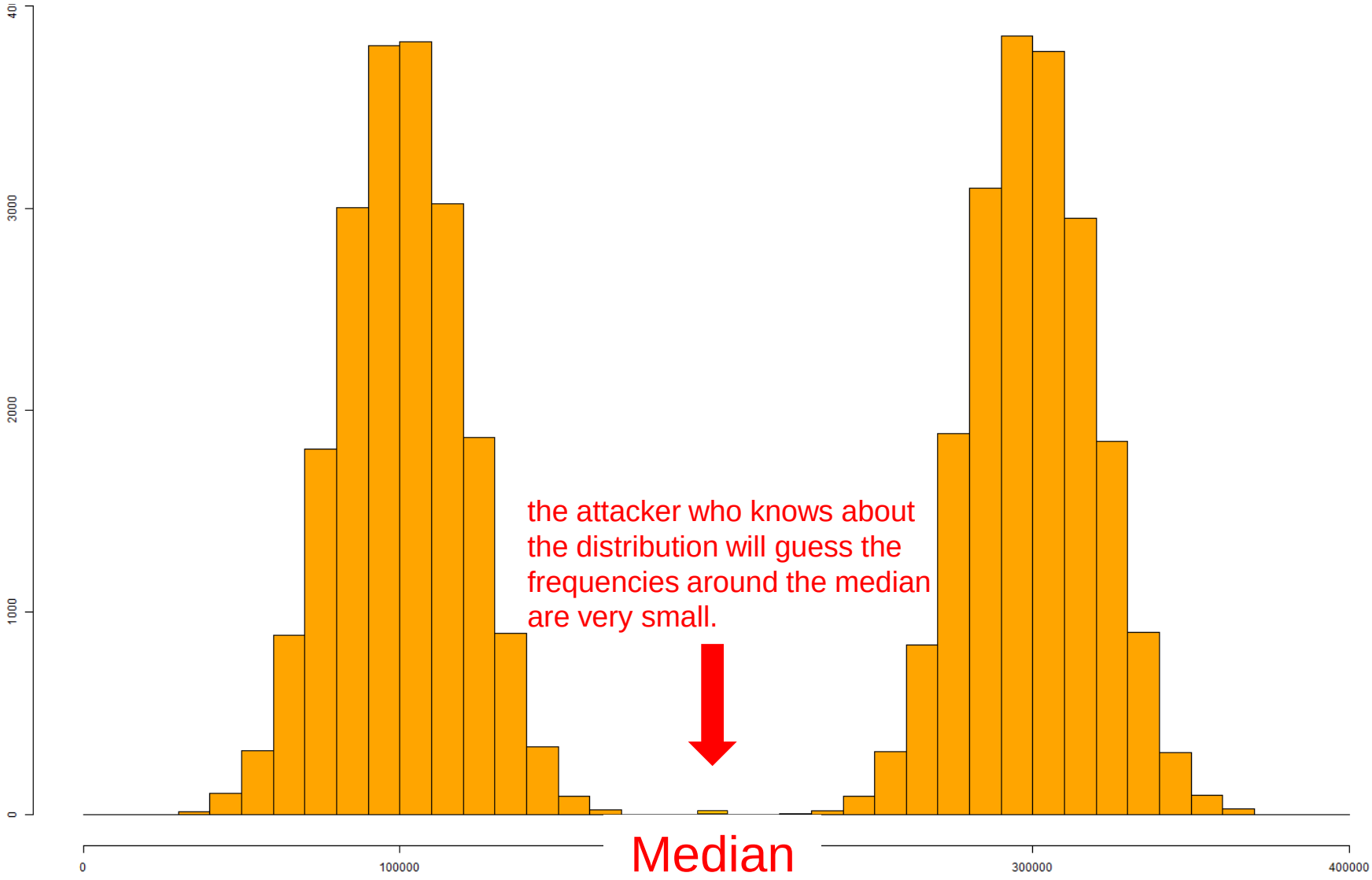
Example of group disclosure on frequency tables by region and income

	0-1 million (yen)	1 million–2 million	2 million–3 million	3 million–	Sum
Region 1	20	20	30	25	95
Region 2	125	5	3	0	133
Region 3	30	30	30	43	133
Sum	175	55	63	68	361

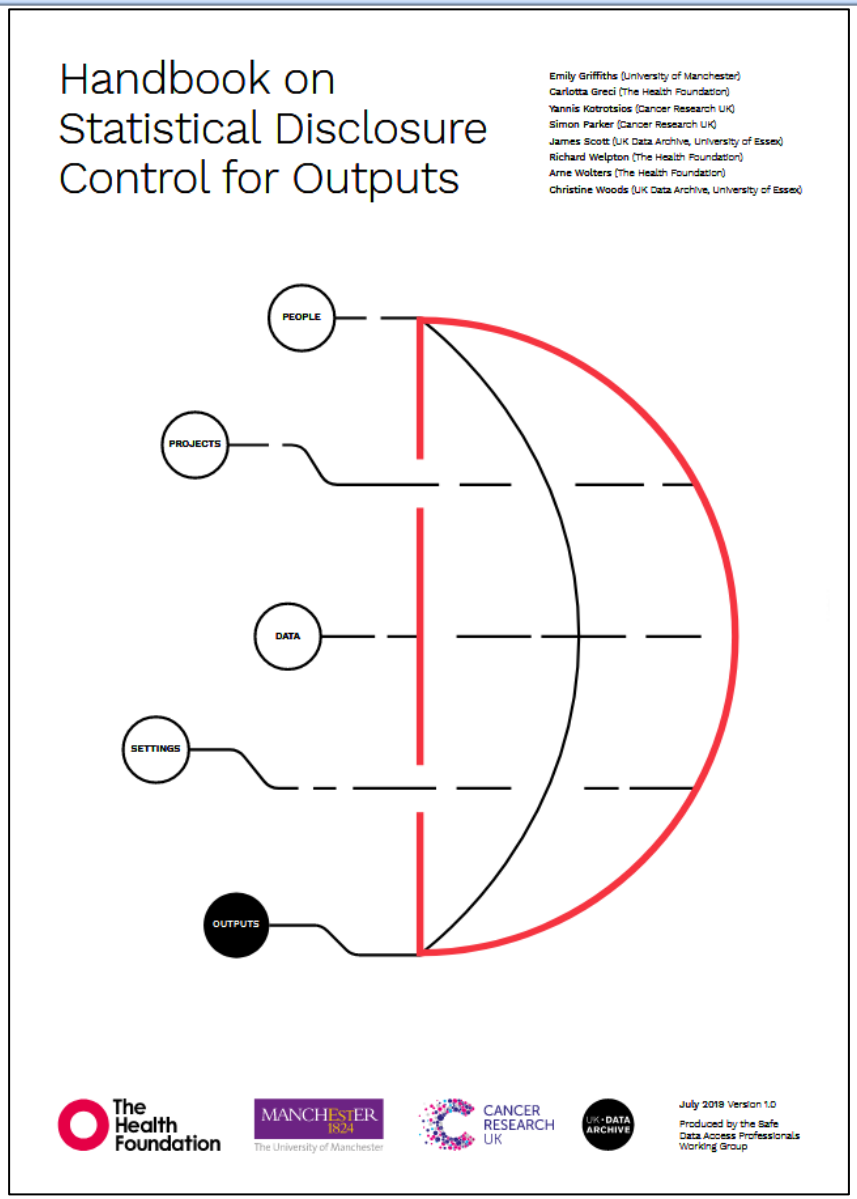
We can estimate income of a resident living in region 2 is 0-1 million with high probability.

Case Studies of DwB Project (5/5)

3. Consider the case where the variance around the percentile is very large.



Case Studies of UK Data Service (1/2)



Handbook on Statistical Disclosure Control for Outputs

Emily Griffiths (University of Manchester)
Carlotta Greci (The Health Foundation)
Yannis Kotrotsios (Cancer Research UK)
Simon Parker (Cancer Research UK)
James Scott (UK Data Archive, University of Essex)
Richard Welpton (The Health Foundation)
Arne Wolters (The Health Foundation)
Christine Woods (UK Data Archive, University of Essex)

July 2019 Version 1.0
Produced by the Safe Data Access Professionals Working Group

UK Data Service (2019, July). Handbook on Statistical Disclosure Control for Outputs [4].

「Rounding Suppression」

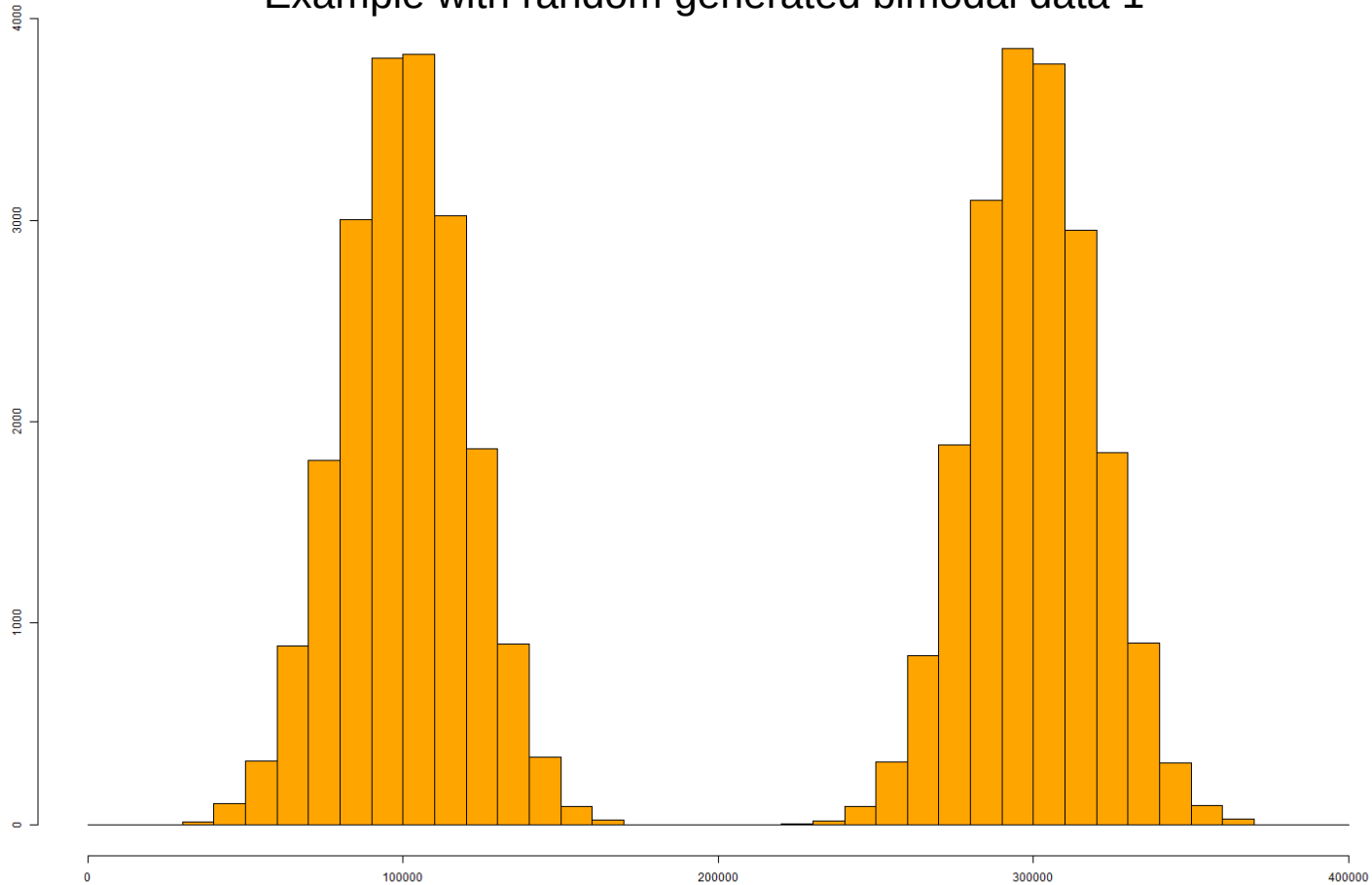
P.33-34

- Increase the number of digits to round the median or quartile value and the every individual value until the frequency of individuals with the same rounded value as the rounded median or the quartile value is 10 or greater.

	1st Quartile	Median	3rd Quartile
True value	3804.9	5503.7	7983.6
Rounded value	3800	5504	7980
Freq. of individuals which has the same rounded value	62	10	35

Experiment of Rounding Suppression (1/3)

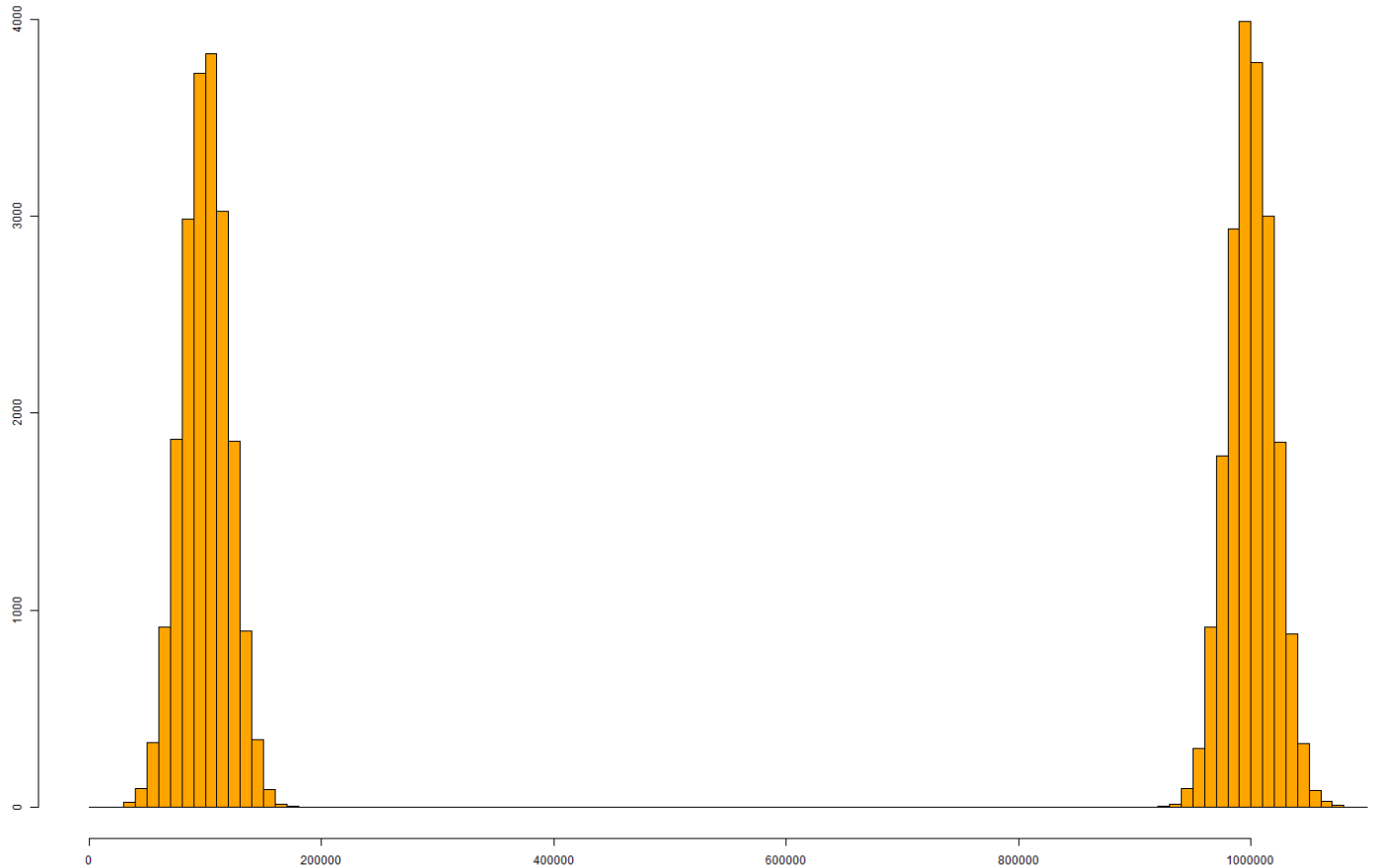
Example with random generated bimodal data 1



	1st Quartile	Median	3rd Quartile
True value	100178.4	196825.0	299731.9
Rounded value	100200	200000	299700
Freq. of individuals which has the same rounded value	32	228	36

Experiment of Rounding Suppression (2/3)

Example with random generated bimodal data 2



	1st Quartile	Median	3rd Quartile
True value	100136.5	550353.1	999868.4
Rounded value	100100	NA	999900
Freq. of individuals which has the same rounded value	39	0	32

Experiment of Rounding Suppression (3/3)

T3. If the variance around the percentile is very large, the identity of the percentile respondent might be guessable.

Solution:

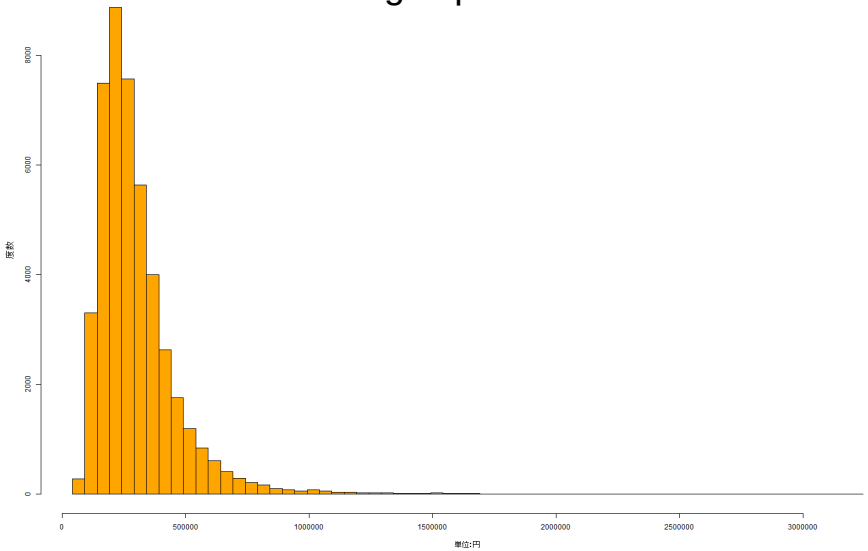
When the variance around the median or quartile is very large, the rounding suppression prevent to publish the value.

Experiment with skewed data (1/2)

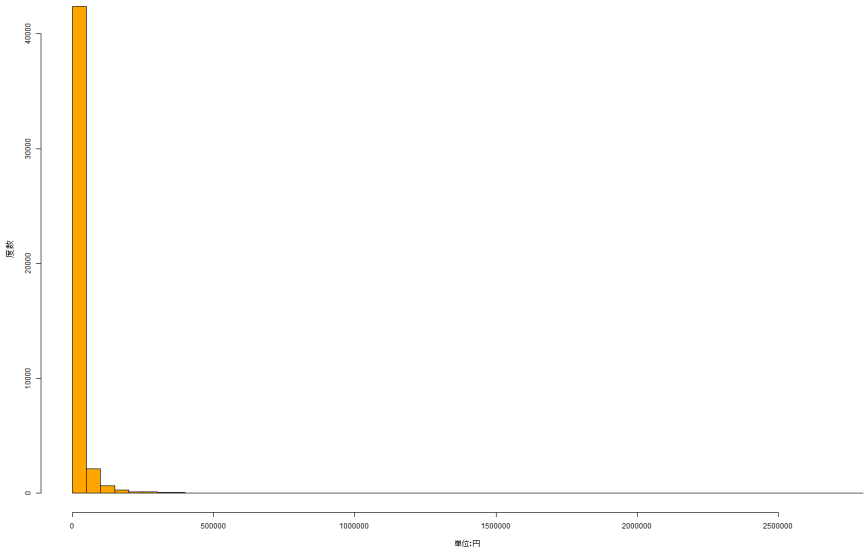
Synthetic data of National Survey of Family Income, Consumption and Wealth 2009
Sample size 45,811

	Mean (yen)	skewness
Yearly income	6401900	2.12
Living expenditure	298373.7	3.01
Food	68740.2	1.26
Housing	16127.9	12.25
Fuel, light and water charges	19421.0	1.36
Furniture and household utensils	9374.0	7.63
Clothes and footwear	12054.8	6.41
Medical care	13281.0	7.17
Transportation and communication	44692.4	7.99
Education	15014.9	14.59
Reading and recreation	31099.3	4.59
Other living expenditure	68568.3	5.94

Living expenditure



Education



Experiment with skewed data (2/2)

	1st Quartile	Rounded 1st Q.	Change rate	Median	Rounded Median	Change rate	3rd Quartile	Rounded 3rd Q.	Change rate
Yearly income	3804.9	3800	0.13%	5503.7	5504	0.01%	7983.6	7980	0.05%
Living expenditure	194216.5	194200	0.01%	260255.8	260300	0.02%	354990.9	355000	0.00%
Food	48104.4	48100	0.01%	63497.8	63500	0.00%	83502.0	83500	0.00%
Housing	659.3	660	0.11%	2876.9	2880	0.11%	17461.6	17500	0.22%
Fuel, light and water charges	13589.7	13590	0.00%	17905.0	17900	0.03%	23544.7	23540	0.02%
Furniture and household utensils	2877.6	2880	0.08%	5613.9	5610	0.07%	11057.8	11060	0.02%
Clothes and footwear	3893.7	3890	0.09%	7686.1	7690	0.05%	14515.6	14500	0.11%
Medical care	3849.0	3850	0.03%	7687.3	7690	0.03%	15434.6	15430	0.03%
Transportation and communication	10372.4	10400	0.27%	23464.7	23460	0.02%	48760.9	48800	0.08%
Education	0.0	0	0.00%	1761.2	1760	0.07%	13535.3	13500	0.26%
Reading and recreation	12182.3	12180	0.02%	21707.7	21710	0.01%	38310.9	38300	0.03%
Other living expenditure	24793.2	24800	0.03%	44634.2	44600	0.08%	80909.8	80900	0.01%

Change rate := | True value - Rounded value | / True value

Frequency Threshold Rule (1/2)

T1. If the rank ordering of firms is known or guessable, the percentile cannot be released.

Our Assumption:

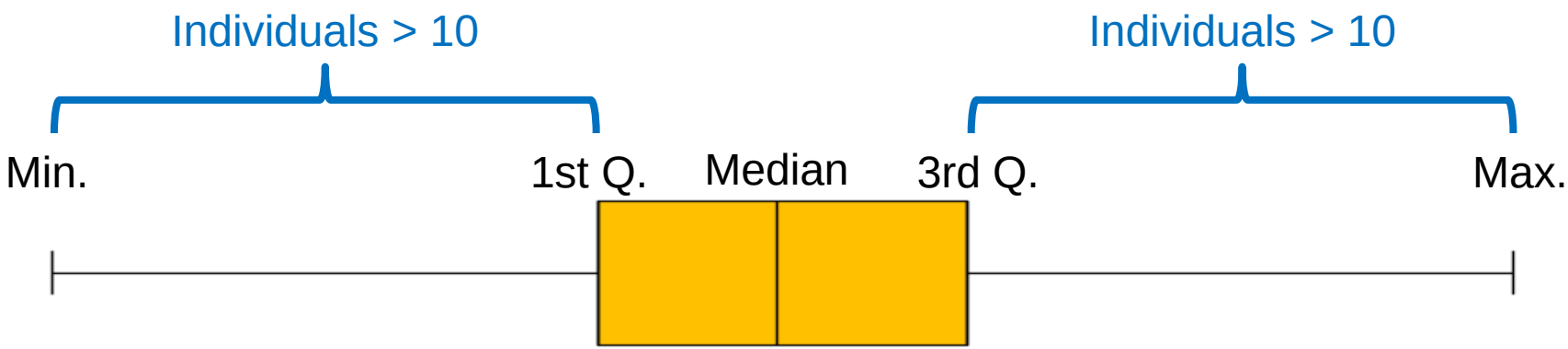
It is possible that the top rankings are known or can be inferred, however in general, if the data size is large enough, it is assumed that it is difficult to accurately determine the rankings of individuals near the median and quartiles.

Our Solution:

Ensure that either 1st or 3rd quartile doesn't belong to the range of the bottom or to rankings.

Frequency Threshold Rule (2/2)

- The frequency of the group for which the median and quartile values are calculated must be at least 40.
  (assume that an attacker can infer top or bottom 10 ranking.)



Interquartile Range Threshold Rule (1/2)

T2. If the variance around the percentile is low, there is the possibility of group disclosure.

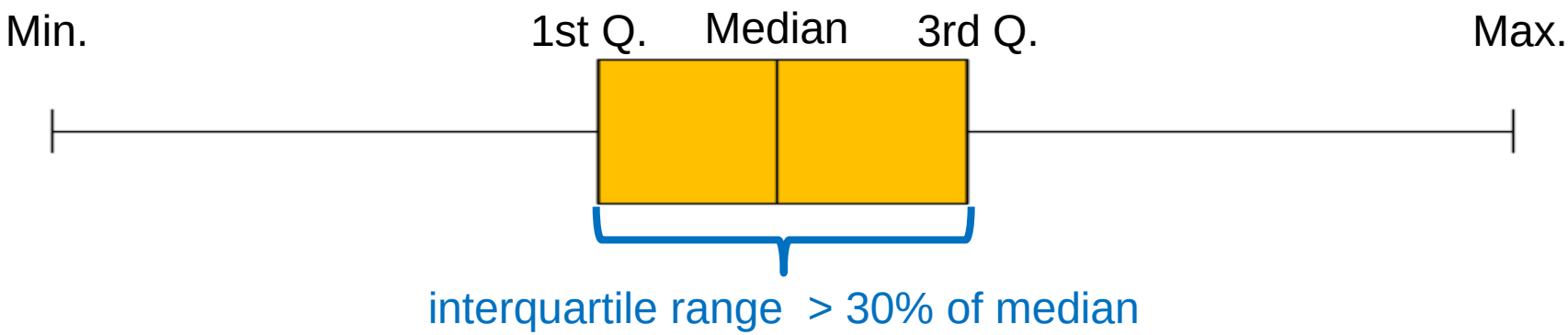
→ We should introduce an additional rule to prevent group disclosure that would apply to sensitive variables as we do for *sum* and *mean*.

Our Solution:

Introduce an additional rule requires that sample has some degree of dispersion.

Interquartile Range Threshold Rule (1/2)

- The interquartile range must be more than 30% of the median.



Conclusion

- The rounding suppression method of UK Data Service is simple yet satisfying the principle of 10 units, and coping with T3 in DwB.
- To cope with T1 and T2 in DwB, we introduce the frequency threshold rule and the additional interquartile range threshold rule.

References

[1] National Statistics Center, "Using microdata of official statistics (in Japanese)," https://www.e-stat.go.jp/microdata/data-use/on-site.

[2] R. Kikuchi and K. Minami, "On-site service and safe output checking in japan," in Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje, North Macedonia, 2017.

[3] Data without Boundaries project, "Guidelines for the checking of output," https://ec.Europa.eu/eurostat/cros/system/files/dwb_standalone-document_output-checking-guidelines.pdf.

[4] UK Data Service, "Handbook on Statistical Disclosure Control for Outputs," https://ukdataservice.ac.uk/app/uploads/thf_datareport_aw_web.pdf.