UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

26-28 September 2023, Wiesbaden

A Case Study of Output Checking in Japan

Yutaka Abe and Kazuhiro Minami (National Statistics Center, Institute of Statistical Mathematics)

yabe3@nstac.go.jp, kminami@ism.ac.jp

Abstract

In Japan, the Japan National Statistics Center has been responsible for checking the output of on-site use of official statistical microdata since its launch in 2019. We have accumulated experience in output checking as we examine how to apply checking rules to various outputs produced by researchers in different research fields. We have also identified the need for adding new rules. In this paper, we present our experiences in output checking as a case study in Japan and describe the new rules for quantiles, which we plan to introduce.

1 Introduction

In 2019, Japanese Statistical Act has been revised and on-site use of official statistical microdata launched. In Japan, microdata of official statistics has traditionally been provided by the media such as DVD. However, because it is not possible to check the outputs created by the users after the media of microdata are provided, the purpose of use and the outputs the users will create are confirmed in advance before the start of use, and only limited variables necessary to create the outputs are provided. In addition, Japan has decentralized statistical system, which mean each ministry has its own statistical survey. Hence the users need to apply to multiple ministries for each statistical survey, which they need for research purposes.

Therefore, for on-site use, we only confirm the outline of the research method and the output image that users will create before the start of use, provides all variables of microdata, and instead, check each output which the users create. Furthermore, each ministry outsources its administration to the National Statistics Center as the central contact point for on-site use procedures. Thus, the National Statistics Center checks all the outputs through on-site use, and is also reviewing the rules for output checking appropriately.

The rest of the paper is organized as follows. In Section 2, we introduce Japanese on-site use institution for microdata of official statistics. In Section 3, we present our experience with output checking as a case study; in Section 4, we discuss a new output checking rule for quartiles that we are considering introducing. Section 5 describes future work.

2 On-site Use for the Microdata of Official Statistics in Japan

2.1 On-site Use Institution in Japan

For the importance of EBPM has been recognized in Japan in recent years, the Statistics Act has been revised in 2019 to make microdata of official statistics available for policymaking and academic research. In Japan, the microdata of official statistics has traditionally been provided only to public institutions and researchers subsidized by public institutions, etc. However, in order to promote further use of the microdata, the requirements for use have been expanded so that, for example, faculty members affiliated with universities in Japan can use the microdata even if they are not subsidized by public institutions.

On the other hand, since the microdata contains many confidential information of the survey individuals, the provision of the microdata by on-site use has also been started in order to provide securely. Especially, for the use requirements added in the 2019 revise are provided only by on-site use.

Since on-site use allows for post-checking of the output, when submitting a using request of the microdata, we confirm the outline of the research method and the output image that users will create. This was intended to shorten the time to start on-site use and to allow for explorative and creative analysis with all variables in on-site use. In addition, the portal site (in Japanese) [1] is open to the public for on-site use procedures.

Japanese output checking rule has been under consideration even before on-site use was officially launched, as presented at the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality held in September 2017 [2]. Following this consideration, the current output checking rules are based on the principles that, each individual value is confidential, calculate statistical values from the individual data values of 10 or more units (principle of 10 units), remain 10 or more degrees of freedom for mathematical models such as regression coefficients (principle of 10 degrees of freedom). Moreover, dominance rule for statistical surveys covering establishments, and optional rules to prevent group disclosure for sensitive variables are determined.

2.2 Overview of On-Site Systems in Japan

Figure 1 shows an overview of Japanese on-site system. As of July 1, 2023, there are 21 on-site facilities in Japan and 19 of which are located at universities. Thin client PCs are established at each on-site facility, and users connect to the central virtual PC server from these thin client PCs to use the survey information on their individual virtual desktops. For secure connection, the communication lines for the on-site system use SINET, an academic backbone network for universities and research institutions built and operated by the National Institute of Informatics, and are disconnected from the general Internet connection.

For security purposes, we keep records of who and when enters and exits the on-site facility, and we monitor and record user activity inside the facility by monitoring cameras. Thin client PCs are set up so that external disks and USB flash drives cannot be used. Therefore, users cannot directly take out the outputs of their research and analysis. We retrieve the outputs from the virtual PC on behalf of the user, check the confidentiality of the outputs and send them to the user if they are safe.

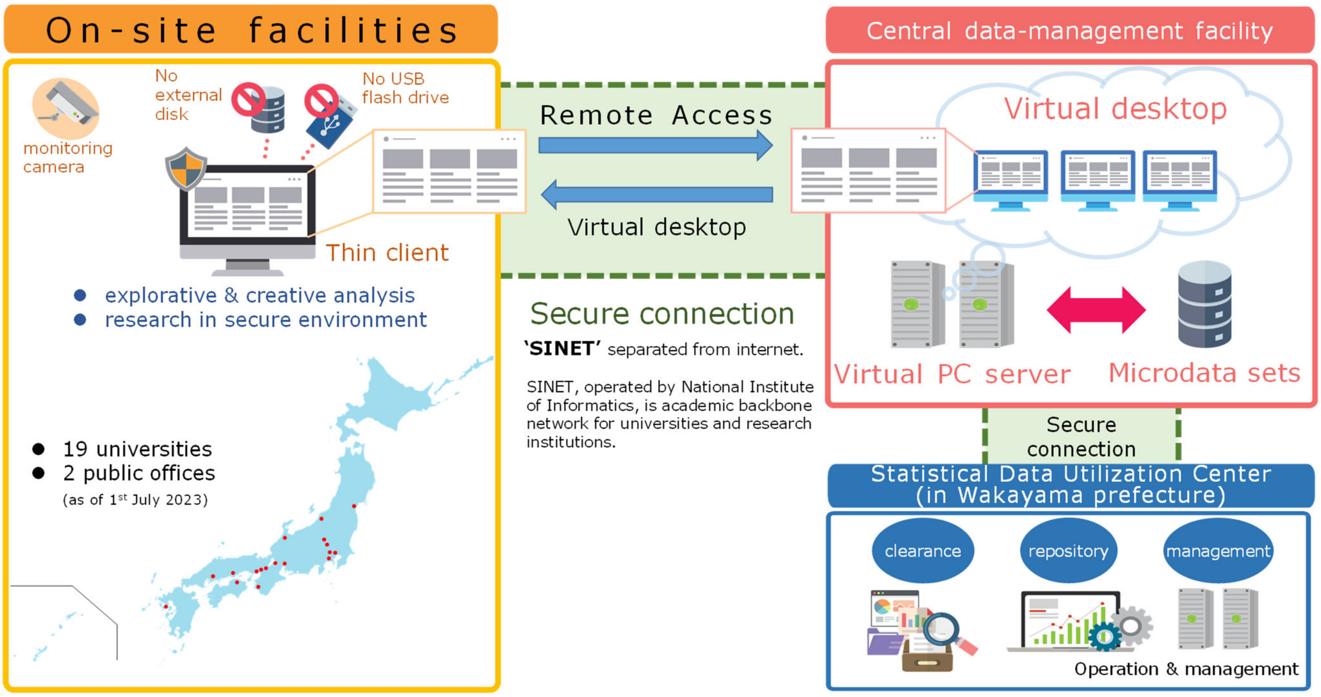
On-site facilities
monitoring camera
No external disk
No USB flash drive
Thin client
● explorative & creative analysis
● research in secure environment
● 19 universities
● 2 public offices
(as of 1st July 2023)
Remote Access
Virtual desktop
Secure connection
'SINET' separated from internet.
SINET, operated by National Institute of Informatics, is academic backbone network for universities and research institutions.
Central data-management facility
Virtual desktop
Virtual PC server
Microdata sets
Secure connection
Statistical Data Utilization Center (in Wakayama prefecture)
clearance
repository
management
Operation & management

Figure 1 Secure microdata access service in On-site facilities

3 A Case Study of Output Checking in Japan

The National Statistics Center checks all of the outputs by on-site use based on predefined output checking rules. The output checking rules are defined for commonly used output formats, however, in the past five years of output checking there were cases that could not check the safety of outputs only by formally applying the rules, and necessary to understand the contents of the outputs and conduct checks other than the rules, or to change the format to enable checks within the rules. In this section, we present a case study of output checking conducted at the National Statistics Center that was suggestive in our work. Note that the values in the case studies are not actual.

3.1 Case of Statistical Tables Containing Hidden Attributes

A user created Table 1 in the on-site facility. The user has analyzed the number of women employed non-regular (part-time, etc.) and created a table of the number of employees by employment status and age for female and male total and for female.

Table 1 Number of employees by gender, employment status and age

Gender	Employment Status	Under 24 y.o.	25-29 y.o.	…
Female & Male	All Status			
Female & Male	Non-Regular			
Female	All Status			
Female	Non-Regular			

However, upon review, it was not sufficient to simply apply the existing rules for frequency table to this table. The difference between total and female we can calculate the male frequencies, and the difference between all status of employment and non-regular employment we can calculate the frequencies of normal employment. As shown in Table 2, since it was necessary to check the statistical tables including the hidden attributes, we asked the user to create a complete table as explanatory materials, and we used the rules of frequency tables to check the safety of the complete table.

Table 2 Statistical tables containing hidden attributes

Gender	Employment Status	Under 24 y.o.	25-29 y.o.	…
Female & Male	All Status			
Female & Male	Regular			
Female & Male	Non-Regular			
Female	All Status			
Female	Regular			
Female	Non-Regular			
Male	All Status			
Male	Regular			
Male	Non-Regular			

3.2 Case of Statistics on Total and Breakdown

A user created Table 3 in the on-site facility. In order to compare the number of manufacturing establishments and other industrial establishments in an area, the user created a table of the establishments' frequency, means and standard deviations of the amount of sales.

Table 3 Statistical tables on manufacturing and other industries (before suppression)

All Industries			Manufacturing			Other Industries		
Freq.	Mean	S.D.	Freq.	Mean	S.D.	Freq.	Mean	S.D.
50	7,074.4	14,373.7	10	13,946.4	31,992.8	40	5,356.4	2,870.9

Our output checking rules specify $(1, 70)$ rule and $(2, 85)$ rule for mean calculated from statistical surveys covering establishments. In this table, since in manufacturing the mean did not satisfy the $(1, 70)$ rule and degrees of freedom of the standard deviation was less than 10, the user suppressed the two corresponding cells as shown in Table 4.

Table 4 Statistical tables on manufacturing and other industries (after suppression by user)

All Industries			Manufacturing			Other Industries		
Freq.	Mean	S.D.	Freq.	Mean	S.D.	Freq.	Mean	S.D.
50	7,074.4	14,373.7	10	X	X	40	5,356.4	2,870.9

However, in the case of Table 5, we can easily recalculate the mean of the manufacturing as follows.

$$\boldsymbol{m_1} = \frac{(n_0 m_0) - (n_2 m_2)}{n_1}$$

We also examined whether the standard deviation could be recalculated, then we found that we can recalculate as follows.

$$\boldsymbol{s_1} = \sqrt{\frac{(n_0 - 1)s_0^2 + 2m_0(n_1 m_1 + n_2 m_2) - n_0 m_0^2 - n_1 m_1^2 - n_2 m_2^2 - (n_2 - 1)s_2^2}{(n_1 - 1)}}$$

Table 5 Recalculation of Mean and Standard Deviation

All Industries			Manufacturing			Other Industries		
Freq.	Mean	S.D.	Freq.	Mean	S.D.	Freq.	Mean	S.D.
n_0	m_0	S_0	n_1	m_1	S_1	n_2	m_2	S_2

Therefore, we coordinated with the user and finally decided to suppress the cells as shown in Table 6 so that they could not be recalculated.

Table 6 Statistical tables on manufacturing and other industries (after final suppression)

All Industries			Manufacturing			Other Industries		
Freq.	Mean	S.D.	Freq.	Mean	S.D.	Freq.	Mean	S.D.
50	7,074.4	14,373.7	X	X	X	X	5,356.4	2,870.9

3.3 Case of Decision Tree

A user created Figure 2 in the on-site facility. This is the decision tree created in R for the conditions under which the employee buys stock. There is no output checking rule for the decision tree, nor in this figure is it possible to check the size of the sample from which it was created or the frequencies at each node. However, we thought that if we could ascertain such information, we can check by applying the existing rules of frequency tables.

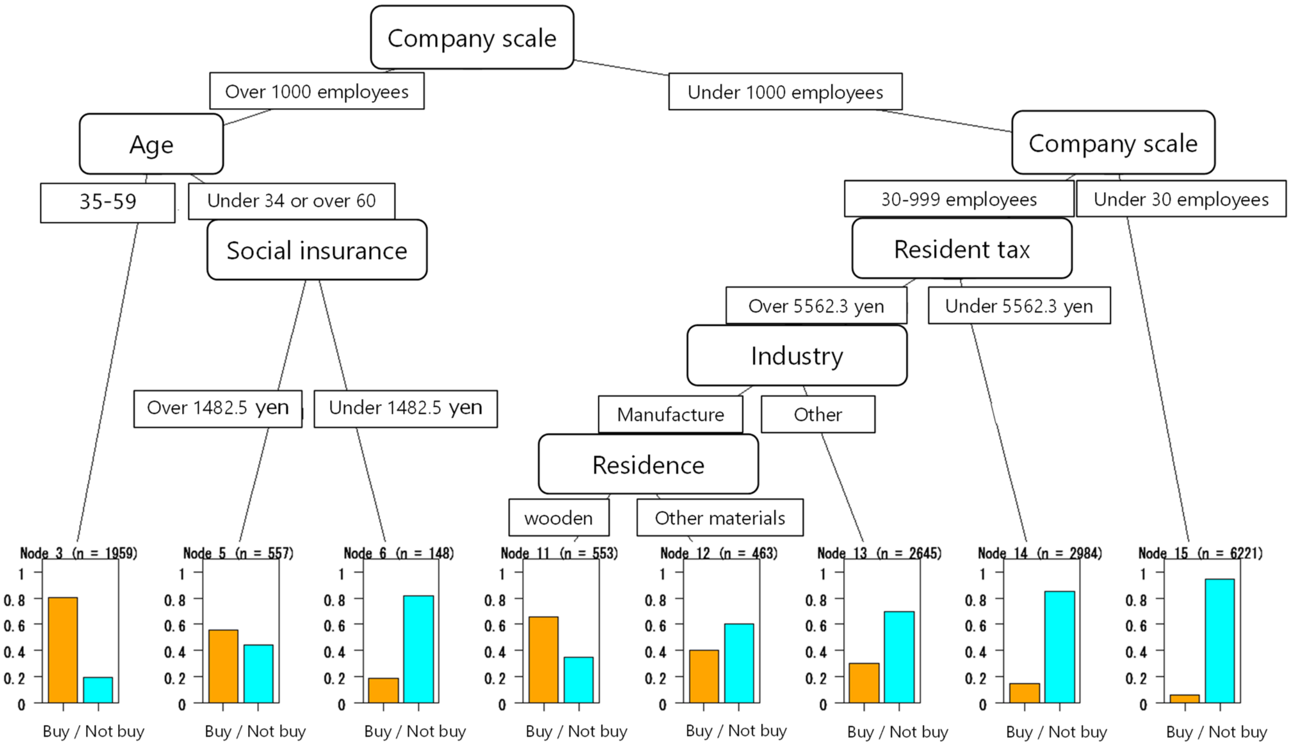

Figure 2 Decision tree regarding whether or not an employee buying stock

Therefore, we considered an R script that converts the structure of the decision tree into a frequency table, and then asked users to create Table 7 and check it using the rules of the frequency table.

Table 7 Conversion of the decision tree into a frequency table

	Buy	Not buy	Total
Total	4033	11497	15530
Company scale over 1000 employees	1917	747	2664
Age 35-59 years old	1579	380	1959
Age under 34 or over 60 years old	338	367	705
Social insurance over 1482.5 yen	311	246	557
Social insurance under 1482.5 yen	27	121	148
Company scale under 1000 employees	2116	10750	12866
Company scale 30-999 employees	1772	4873	6645
Resident tax over 5562.3 yen	1341	2320	3661
Industry manufacture	547	469	1016
Residence wooden	362	191	553
Residence other materials	185	278	463
Industry other	794	1851	2645
Resident tax under 5562.3 yen	431	2553	2984
Company scale under 30 employees	344	5877	6221

4 Consideration of Output Checking Rules for Quantiles

4.1 Previous Case Studies

Since the Japanese statistical system is decentralized, each ministry owns their microdata of official statistics. The National Statistics Center is the central contact point for on-site services, however, if there are no predetermined output checking rules for an output that user created, we need to discuss with ministries that own the microdata regarding the checking method, and this requires time to provide the output. Therefore, for commonly used output formats we should determine output checking rule in advance to reduce the time required for provide.

The current output-checking rules require that, in principle, create outputs from 10 or more units. The median and quartiles are widely used in descriptive statistics, etc., however, they do not satisfy the principle of 10 units because their values are either the values of an individual or values obtained from two individuals' values. On the other hand, since it is generally difficult to accurately know the rankings of all survey individuals, we thought it would be possible to establish median and quartile rules by setting appropriate rules, and first collected information on previous cases.

4.1.1 Case Studies of Data without Boundaries Project

The guideline of Data without Boundaries Project [3] stated the following principles for checking percentiles.

1. If the rank ordering of firms is known or guessable, the percentile cannot be released.
2. If the variance around the percentile is low, there is the possibility of group disclosure.
3. If the variance around the percentile is very large, the identity of the percentile respondent might be guessable

Regarding 1, 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.
Regarding 2, consider establishing an optional rule to prevent group disclosure that would apply to sensitive variables.
Regarding 3, for example as shown in Figure 3, if the frequencies around the median are small, there is a risk that if the attacker has knowledge of the distribution, the attacker will guess that the frequencies around the median are very small, and it is necessary to establish rules to prevent this.

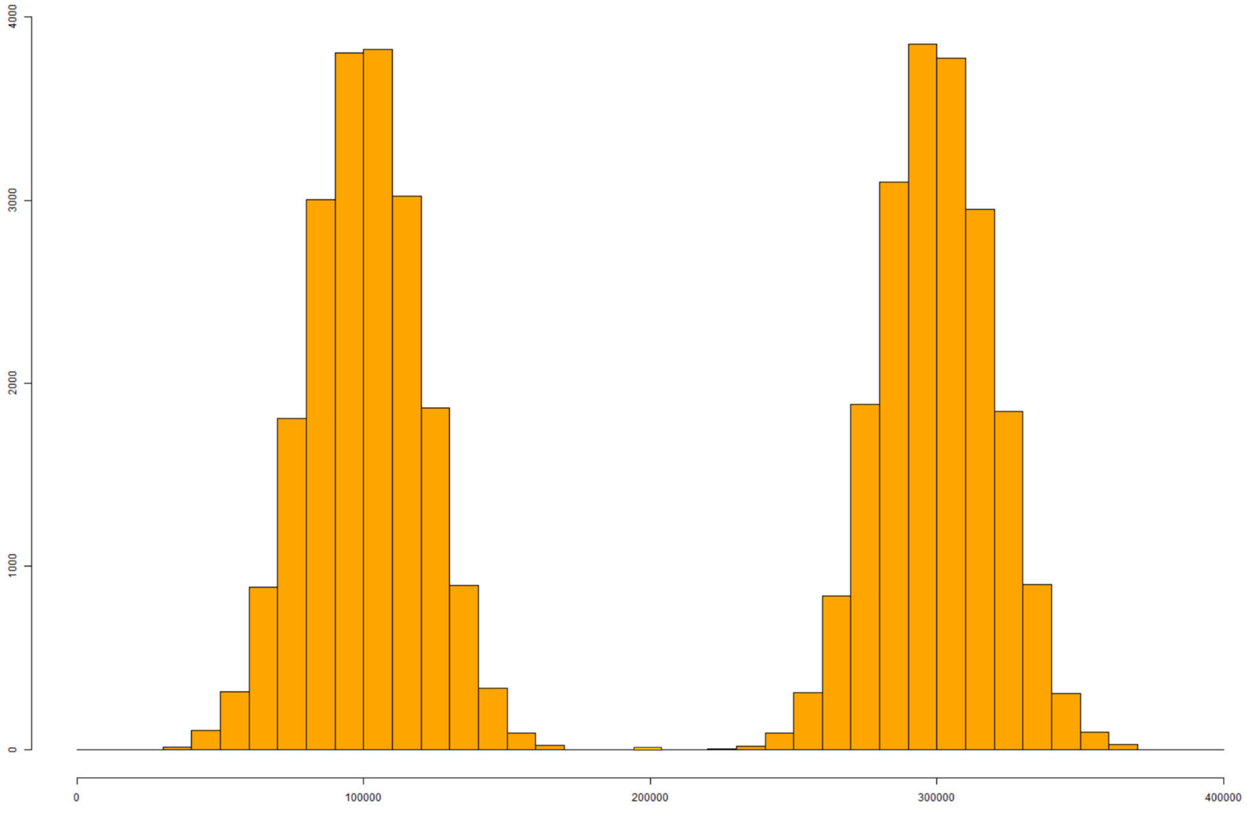

Figure 3 Bimodal distribution with a small number of values around the median

4.1.2 Case Studies of UK Data Service

The Handbook UK Data Service [4] introduce the rounding suppression method. This is a method of increasing 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. For example, Table 8 shows that if the median and every individual value are rounded to one decimal place, there are 10 units with the same rounded value as the rounded median. For the first and third quartile values, rounding off to the nearest ten, there are 10 or more units with the same rounded value as the rounded quartiles.

Table 8 Example of rounding suppression

	first quartile	median	third quartile
True values	3804.9	5503.7	7983.6
Rounded value	3800	5504	7980
Freq. of individuals which has the same rounded value	62	10	35

The verification of this suppression method showed that if the variance around the median is very large, even if the number of rounding digits is increased up to the upper limit, the frequencies of the same rounded values are not more than 10, and the median not provided; otherwise, it is possible to conceal it, although there is variation in the number of rounding digits.

4.2 Results of the Consideration

Further consideration with the results of the information gathering, we are considering that setting two rules in addition to the UK Data Service's rounding suppression method.
One is the rule that the frequency of the group for which the median and quartile values are calculated must be at least 40, in order to ensure that the frequency of survey individuals with values below the first quartile or above the third quartile is at least 10.
The other is an optional rule for sensitive variables that requires some degree of dispersion around the median to prevent group disclosure. Specifically, we consider the rule that the interquartile range, which mean the difference between the third quartile and the first quartile, must be more than 30% of the median.

5 Future Work

On-site use in Japan began in 2019. In Japan, on-site use of the microdata of official statistics over the past five years has been mainly in the fields of economics and sociology, although use in other fields, such as medical research, has been increasing. With the increase in the use of new fields in the future, it is expected that the number of outputs in new formats will increase, and there is also concern about the creation of outputs whose contents are extremely difficult to understand, such as machine learning. Therefore, we continue to record cases in our operations and will review the output checking rules appropriately based on case studies of other countries.

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.