Optimal Boundary Value for Creating Anonymized Microdata: Empirical Analysis based on Economic Survey Data

Kiyomi Shirakawa, Hitotsubashi University / National Statistics Center
Ryota Chiba, Hitotsubashi University
Yutaka Abe, National Statistics Center

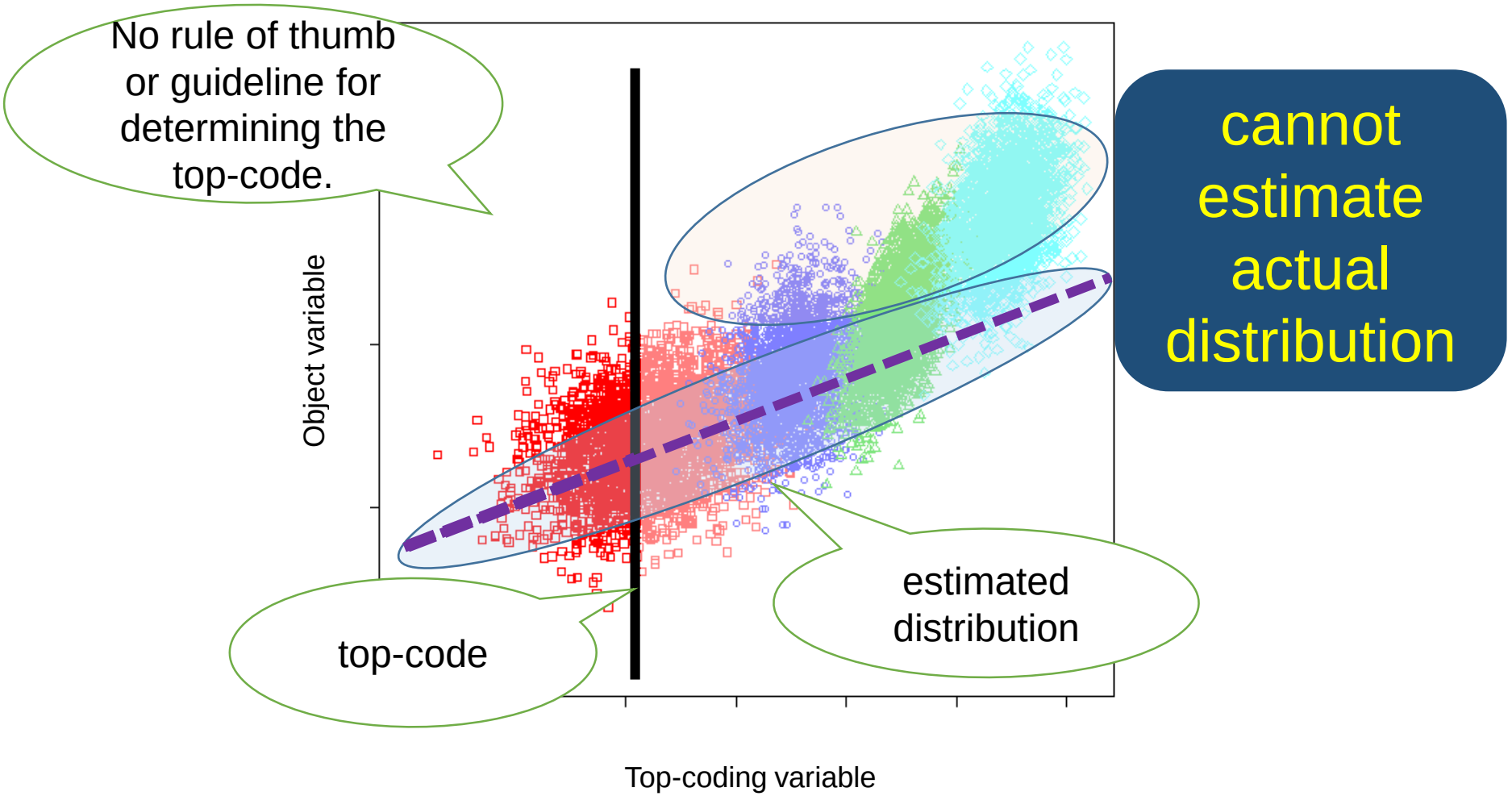
September 13, 2018
Use of R in Official Statistics 2018

Top coding

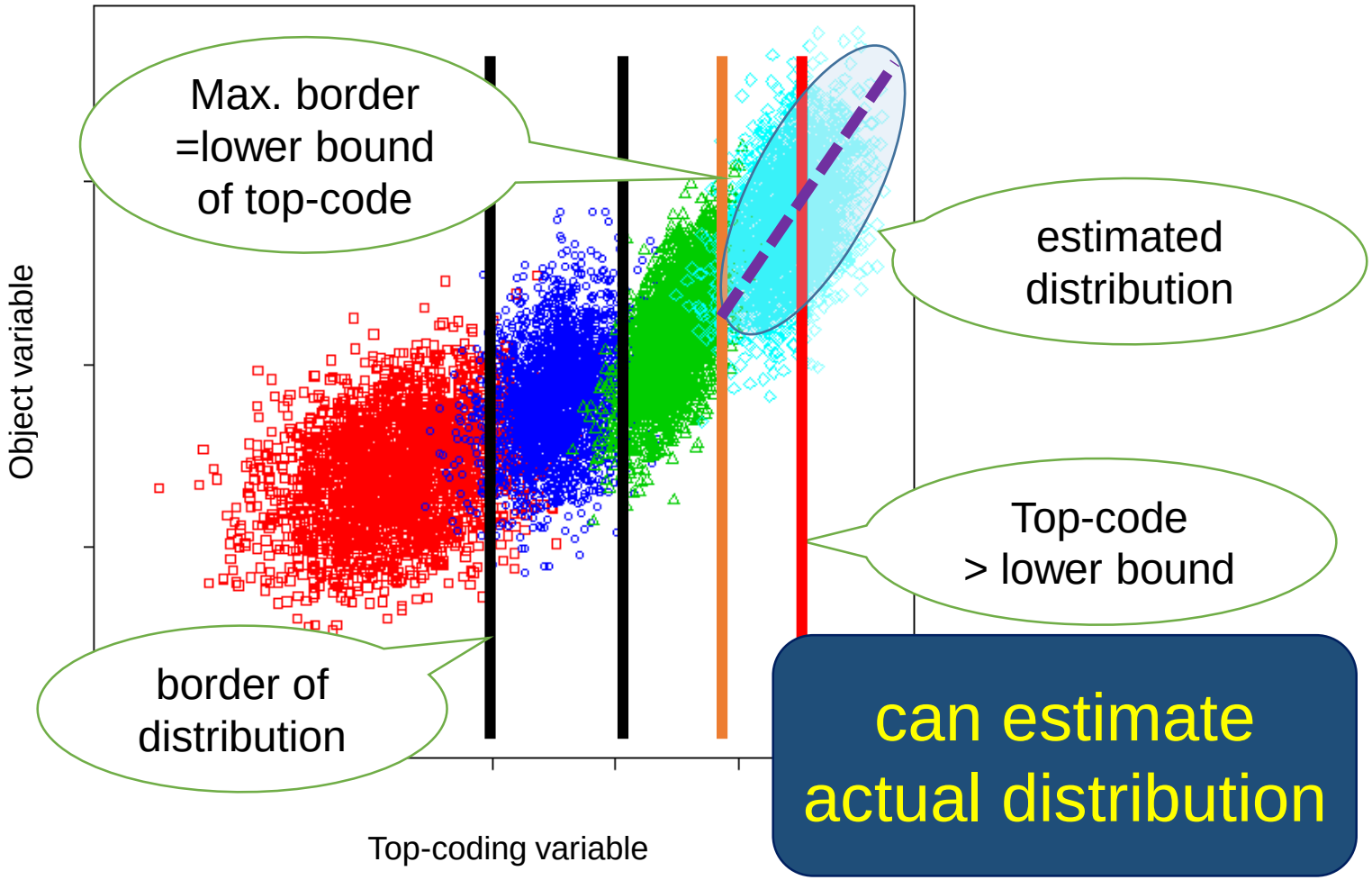
- A method to anonymize quantitative variables.
- A top-code: an upper limit on values of that variable.
- Any value greater than the top-code is replaced by same value or category.

- Example:
  Top-coding variable: family income
  Top-code: 25 million yen
  Family income value greater than 25 million yen is replaced by “over 25 million yen”

Problem on top-coding of official statistics in Japan



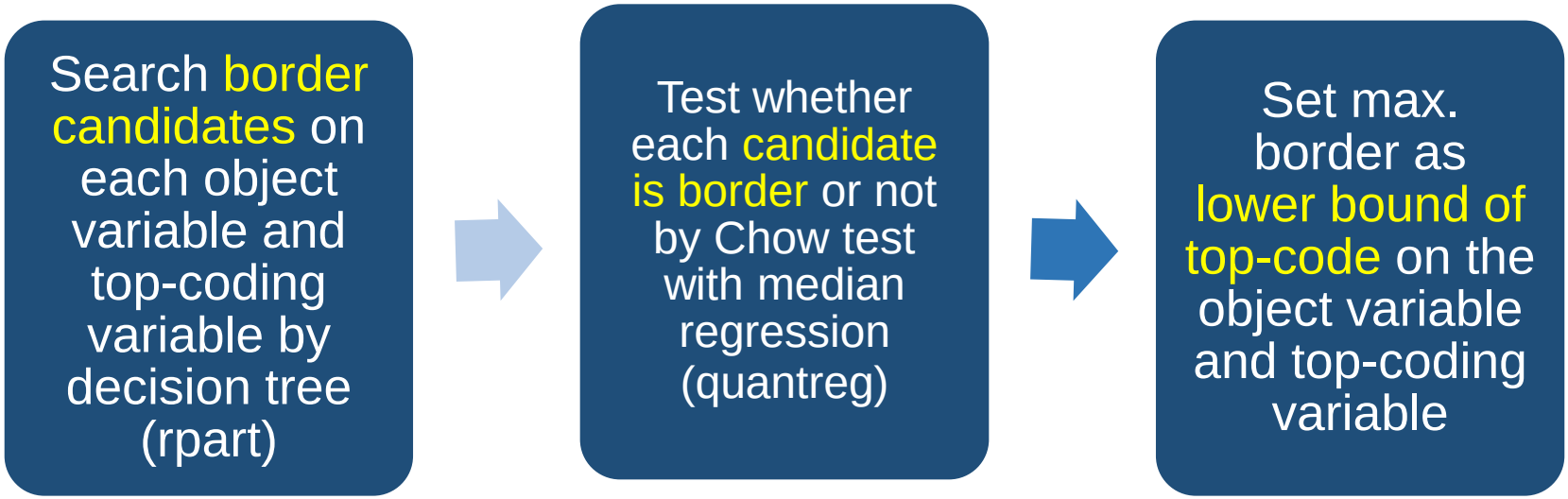
Our idea


Max. border
=lower bound
of top-code
estimated
distribution
Top-code
> lower bound
border of
distribution
can estimate
actual distribution
Object variable
Top-coding variable

R program to estimate lower bound (1)

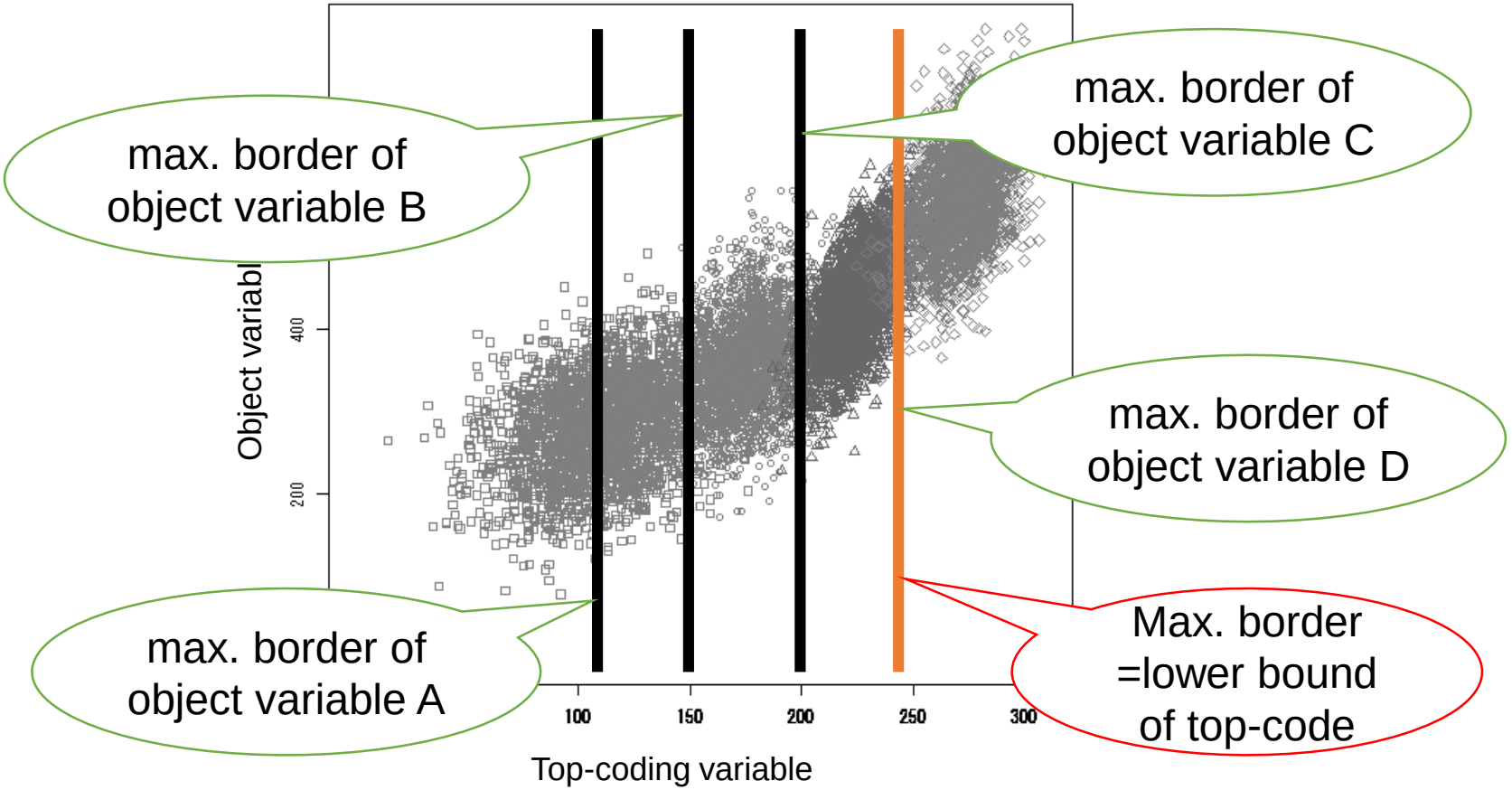
Input: top-coding variable and some object variables

For each object variable,



Then, select max. lower bound of all lower bounds,
and set it as **lower bound of top-code on top-coding variable**.

R program to estimate lower bound (2)



Experiment using actual data

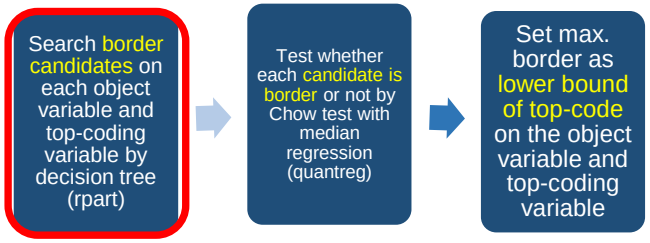
- Dataset: questionnaire information of “2015 Survey of Research and Development”.
  - Survey to provide basic materials for promoting science and technology in Japan, by studying the research and development (R&D) activities carried out in Japan.
- Data size (after data-cleaning): 4,759
- Top-coding variable: Sales
- All variables used in this experiment are log-transformed.

Object variables selection

- Select some variables with high correlation with sales as objective variables.

log transformed	# employees	capital	expenditure on R&D	labor costs	# employees in R&D	# researchers	# main researchers
Sales	0.924	0.783	0.705	0.678	0.672	0.672	0.596

Search border candidates by decision tree (1)

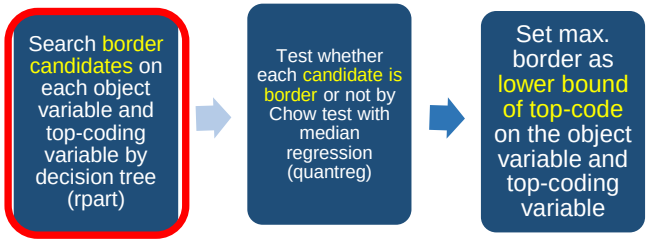


- Use rpart package.

rpart(*object variable* ~ Sales, data=data.log, method="anova") $splits[, 4]

log transformed	Border Candidate 1	Border Candidate 2	Border Candidate 3	Border Candidate 4	Border Candidate 5	Border Candidate 6	Border Candidate 7
# employees	5.139	6.717	7.924	8.849	9.997	11.280	12.381
capital	8.121	8.827	9.672	10.281	11.431	12.436	-
expenditure on R&D	7.182	9.159	10.523	11.615	13.074	-	-
labor costs	4.127	7.245	9.151	10.377	11.615	13.074	-
# employees in R&D	7.185	8.609	10.377	12.387	-	-	-
# researchers	7.185	8.639	10.377	12.434	-	-	-
# main researchers	8.434	10.329	12.387	-	-	-	-

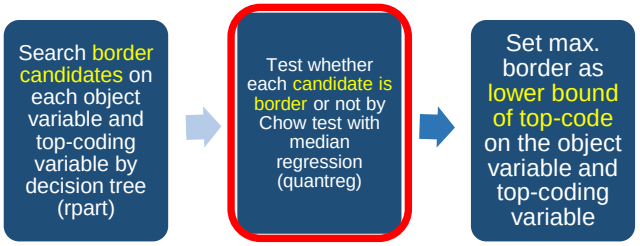
Search border candidates by decision tree (2)



- Divide data by borders on each object variable.

log transformed	Devided data size 1	Devided data size 2	Devided data size 3	Devided data size 4	Devided data size 5	Devided data size 6	Devided data size 7	Devided data size 8
# employees	223	448	684	731	1,109	919	401	244
capital	1,493	578	801	571	752	334	230	-
expenditure on R&D	888	1,476	1,260	647	364	124	-	-
labor costs	126	802	1,429	1,160	754	364	124	-
# employees in R&D	889	986	1,642	1,001	241	-	-	-
# researchers	889	1,015	1,613	1,010	232	-	-	-
# main researchers	1,741	1,737	1,040	241	-	-	-	-

Test each candidate by Chow test with median regression



- Use quantreg package.

rq(*object variable* ~ Sales, data=data.log, tau=0.5)

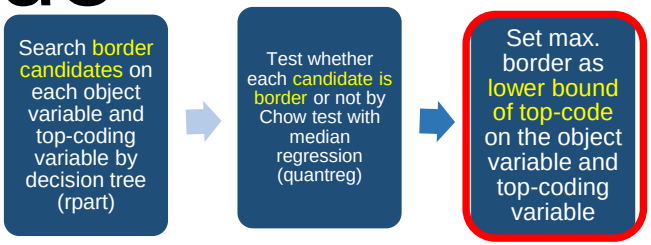
http://aoki2.si.gunma-u.ac.jp/R/Chow.html

# employees (log)	Border Candidate	p-value of Chow test
Border Candidate 1	**5.139**	**0.000**
Border Candidate 2	6.717	0.514
Border Candidate 3	7.924	0.229
Border Candidate 4	8.849	0.662
Border Candidate 5	9.997	0.339
Border Candidate 6	11.280	0.874
Border Candidate 7	12.381	0.467

capital (log)	Border Candidate	p-value of Chow test
Border Candidate 1	**8.121**	**0.000**
Border Candidate 2	8.827	0.088
Border Candidate 3	**9.672**	**0.005**
Border Candidate 4	**10.281**	**0.043**
Border Candidate 5	11.431	0.073
Border Candidate 6	12.436	0.802

$p < 0.05$

Lower bound of top-code

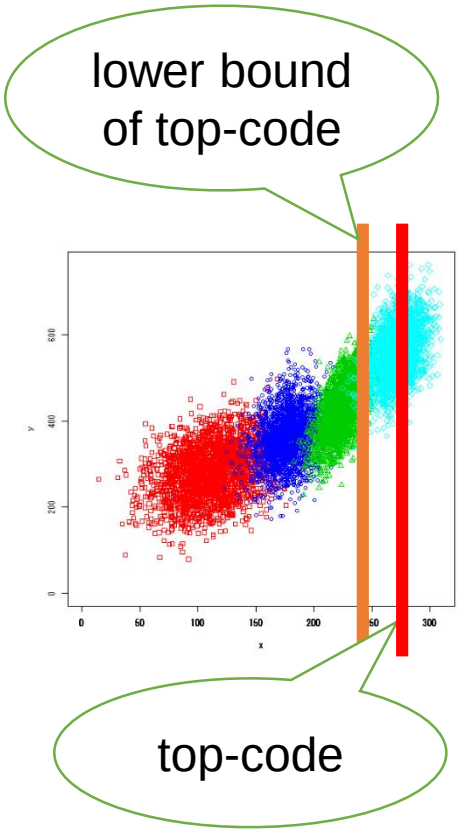


- Select max. lower bound of all object variables.

log transformed	Border
# employees	5.139
capital	10.281
expenditure on R&D	13.074
labor costs	13.074
# employees in R&D	12.387
# researchers	12.434
# main researchers	8.434
lower bound of sales top-code	13.074

Result

Data size	**4,759**
Lower bound of sales top-code (log)	**13.074**
Lower bound of sales top-code	**476,243**
Freq. over lower bound	**124**
% over lower bound	**2.6%**



Half of the distribution: 1.2 ~ 1.3%

Example of top-code	**1,000,000**
Example of top-code (log)	**13.816**
Freq. over top-code	**57**
% over top-code	**1.2%**

Test utility of the top-coding (1)

- In distribution greater than lower bound, test whether
  - ➢ distribution less than top-code and
  - ➢ distribution greater than top-code

  belong same distribution or not, by Chow test with median regression on each object variables.

Test utility of the top-coding (2)

log transformed	p-value of Chow test
# employees	1.000
capital	1.000
expenditure on R&D	1.000
labor costs	0.999
# employees in R&D	1.000
# researchers	0.999
# main researchers	1.000

Conclusion and future development

- We showed an example of rule to decide the top-code from the viewpoint of usability.

- We will verify the top-coding data by this method from the viewpoint of confidentiality.
- We would like to verify how the lower bounds will be changed depending on the year of the survey (if we can get time-series data…).