gesis Leibniz Institute for the Social Sciences

When to use the *k*-rule?

Anja Perry (she/her)

IASSIST, Philadelphia, PA, June 2nd, 2023

Leibniz Association

"As open as possible, as closed as necessary"

(European Commission, 2016)

Anonymization strategies:

- Delete direct identifiers
- Identify quasi-identifiers:
  - Delete
  - Aggregate
  - Top and bottom code

→ Underlying population gets larger

But what about the *k*-rule?

k-anonymity and *l*-diversity

Definition

- No fewer than a certain number (*k*) of individuals, with same indirect identifiers (*k*-anonymity)
- This group cannot have the same characteristics (*l*-diversity)

Advantage

- Clear and transparent rule
- Criteria to determine whether data is anonymized

Disadvantage

- Very rigid rule
- Information loss, especially in high-dimensional data

Used for full censuses and for very visible individuals (e.g., politicians, figures in the public eye)

(Aggarwal, 2005; El Emam and Dankar, 2008; Ritchie and Elliot, 2015)

When should we apply *k*-anonymity?

Sensitivity and the risk assessment matrix

- Art. 9 GDPR
- Further information, such as test results, opinion about employer, illegal actions, …

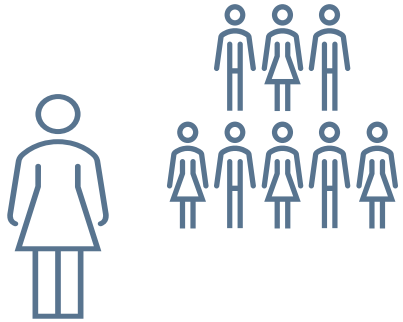
		Data situation sensitivity		
		Low	Medium	High
Summary risk	High	Essential	Essential	Essential
	Medium	Borderline	Essential	Essential
	Negligible	Unnecessary	Borderline	Borderline

Elliot et al. (2020), p.68

The problem of uniqueness

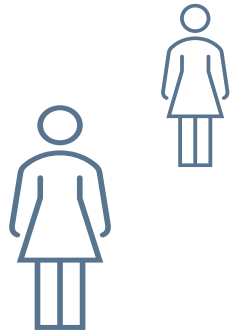
Sample uniqueness

- Respondents who do not share the same combination of characteristics with anyone else
- The smaller the sample and the larger the population, the less critical



Confidence in population uniqueness

- Higher in small populations and when coverage is high
- Also critical: Very visible persons with additional information publicly available (Skinner et al., 1994)

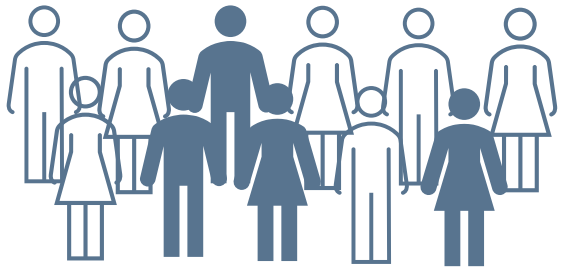


Müller, Blien and Wirth (1995)

Factors increasing confidence in population uniqueness

Representativeness

- Highest when every person is surveyed
- Lower in sampled surveys, but can be high for certain subgroups



Compatibility

- Information in the survey must be compatible with information at hand
- High for geographical information
- Low/non-existent for attitudes, values, …

Müller, Blien and Wirth (1995)

Criteria for *k*-anonymity

		Sensitivity	
		low	**high**
Representativeness and Compatibility	**high**		**k-rule**
	low		

Application to use cases

Disclaimer:

This work was discussed with a member of the Ethical Board at GESIS: Efforts to improve anonymization and protection of respondents

Five use cases

General population survey – Eurobarometer 92.3

- **Low** representativeness
- **Low** to **high** compatibility
- ***k*-rule not necessary**

German Party Membership Study 2017

- **Low** representativeness
- **Low** to **high** compatibility, potential public figures
- ***k*-rule not necessary**

Cologne Dwelling Panel

- **High** representativeness
- **High** compatibility
- ***k*-rule is advised**

Data Sharing Behavior of Researchers

- **High** representativeness
- **High** compatibility
- ***k*-rule is advised**

EU LGBTI Survey 2019

- **Low** representativeness
- **Low** compatibility
- ***k*-rule not necessary**

(**Sensitivity** considered **high** for all use cases.)

General population sample – Eurobarometer 92.3

		Sensitivity	
		Low	High
Representativeness	High		
	Low		

		Sensitivity	
		Low	High
Compatibility	High		
	Low		

- Sensitivity information about political attitudes (Art. 9 GDPR)
- Low representativeness
- Mostly low compatibility, but some countries with detailed regional information

→ **k-rule not necessary**

Political party members – German Party Membership Study 2017

		Sensitivity	
		Low	High
Representativeness	High		
	Low		

		Sensitivity	
		Low	High
Compatibility	High		
	Low		

- Sensitive information about political attitudes and voting decisions (Art. 9 GDPR)
- Low representativeness
- Usually low compatibility
  - EXCEPT: person in the public eye

→ **k-rule not necessary**

Geographically restricted area – Cologne Dwelling Panel

		Sensitivity	
		Low	High
Representativeness	High		
	Low		

		Sensitivity	
		Low	High
Compatibility	High		
	Low		

- Sensitive information, f.ex. also about same-sex relationships (Art. 9 GDPR)
- Highly representative due to small regional coverage
- Highly compatible information especially due to panel design
- We cannot rule out participation knowledge in this small setting

→ **k-rule is advised**

Known and visible sample –
Data sharing behaviour of researchers in sociology and political science

		Sensitivity	
		Low	High
Representativeness	High		
	Low		

		Sensitivity	
		Low	High
Compatibility	High		
	Low		

- Information about religion (Art. 9 GDPR) and data sharing behaviour
- Highly representative, sample can be recreated based on published article
- Highly compatible information as CVs often publicly available

→ **k-rule is advised**

Contributions, limitations and further work

- Check routine and criteria for applying *k*-anonymity
- But no clear rule, only possible criteria
  - Thresholds for sensitivity, representativeness, and compatibility unclear
  - No recommendation for optimal *k* (typically 3 or 5, Thompson and Sullivan, 2020)
- Attention needs to be paid to *l*-diversity!
- Problem with future panel waves: original *k* may become obsolete
- Apply to further datasets

Thank you!

gesis Leibniz Institute for the Social Sciences

Leibniz Association

Contact

Dr. Anja Perry

anja.perry@gesis.org

Tel: +49 221 47694-464

@Datendealerin@fediscience.org

References

- Aggarwal, C. C. 2005. 'On K-Anonymity and the Curse of Dimensionality'. in *Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005*. Trondheim, Norway.
- El Emam, K., and F. K. Dankar. 2008. 'Protecting Privacy Using K-Anonymity'. *Journal of the American Medical Informatics Association* 15(5):627–37. doi: 10.1197/jamia.M2716.
- Elliot, M. E. Mackey, and K. O'Hara. 2020. *The Anonymisation Decision-Making Framework 2nd Edition: European Practitioners' Guide*. Manchester: UKAN.
- European Commission. 2016. H2020 Programme - Guidelines on FAIR Data Management in Horizon 2020. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
- European Parliament, and Council of the European Union. 2018. *General Data Protection Regulation 2016/678*.
- Müller, W., U. Blien, and H. Wirth. 1995. 'Identification Risks of Microdata'. *Sociological Methods and Research* 24(2).
- Ritchie, F., and M. Elliot. 2015. 'Principles- Versus Rules-Based Output Statistical Disclosure Control In Remote Access Environments'. *IASSIST Quarterly* 39(2):5. doi: 10.29173/iq778.
- Skinner, C., C. Marsh, S. Openshaw, & C. Wymer (1994). Disclosure control for census microdata. *Journal of Official Statistics*, 10(1).
- Thompson, K. A., C. & Sullivan (2020). Mathematics, risk, and messy survey data. *IASSIST Quarterly*, 44(4). doi: 10.29173/iq979.