

Wissenschaftliches Arbeiten und Forschungsmethoden

Einheit 11: Publikation wissenschaftlicher Daten

11.01.2024 | Dr. Caroline Zygar-Hoffmann

Heutige Themen

Lehrevaluation

Open Data

Übung: Wie einfach ist Deanonymisierung

Anonymisierungs-Möglichkeiten

Praxis

Lehrevaluation

"Die Studierenden finden die jeweils individuell freigeschalteten Evaluationen im studynet"

The graphic features a collage background showing hands writing on paper. Overlaid text includes "GUTE LEHRE FÖRDERN! Evaluation ist JETZT." and "DEINE LEHREVALUATION im Wintersemester 2023/24". It also includes icons representing efficiency, professionalism, speed, flexibility, and global reach, along with a QR code and the URL "hsf.click/Eval-CFH". Logos for "evasys" and "CHARLOTTE FRESENIUS HOCHSCHULE" are present.

GUTE LEHRE FÖRDERN!
Evaluation ist JETZT.

DEINE LEHREVALUATION
im Wintersemester 2023/24

Im studynet: hsf.click/Eval-CFH

evasys

Jetzt ausfüllen!

CHARLOTTE FRESENIUS
HOCHSCHULE
UNIVERSITY OF PSYCHOLOGY

hsf.click/Eval-CFH

Open Data

Wieso offene Daten?

- Kernkriterien für Wissenschaft: Transparenz & Reproduzierbarkeit (Lupia & Elman, 2014; Merton, 1942; Popper, 1959)
- "Quality Check"
- Nachnutzung: Mehr Erkenntnisse aus den Daten möglich!
- Daten als Grundlage für Meta Analysen , v.a. bei nicht publizierten Studien
- Deutsche Forschungsgemeinschaft (DFG): "Sofern die Daten in Projekten erarbeitet wurden, die aus öffentlich rechtlichen Mitteln finanziert wurden, stehen sie im Grundsatz der Öffentlichkeit frei zur Verfügung" → Wissenschaftler werden von der Gesellschaft finanziert!
- in gewissem Sinne auch: Backup der Daten

Datendokumentation

- Um offene Daten verständlich und dadurch gut nutzbar zu machen braucht es eine gute Datendokumentation → Codebook
- Im Codebook wird für jede Variable beschrieben ...
 - ... wie sie heißt (idealerweise werden dennoch Variablennamen so selbsterklärend wie möglich gewählt)
 - ... was für einen Variablentyp (kontinuierlich/diskret) sie hat
 - ... welche Instruktion sie hatte
 - ... welche Antwortmöglichkeiten es gab
 - ... wie diese Antwortmöglichkeiten kodiert wurden

Open Data

Datendokumentation

Weiterführende Informationen sind immer sinnvoll, z.B. Quellenangabe zu den genutzten Fragebögen, oder falls relevant Erklärung zu fehlenden Werten/Werten die nicht Teil der Antwortoption sind:

Information	Description	Individual item	Scale score	Demographic	Technical
Name of the variable	The name of the variable, exactly as it is displayed in the dataset	BFI_extra_1	CFT-total	Gender	Time
Type of the variable	The type of the variable, usually numeric, factor, date, or character	Numeric	Numeric	Factor	Date
For questionnaire items					
Wording of the item	The original wording of the item, in its original language	„Ich gehe aus mir heraus, bin gesellig.“	–	„Bitte geben Sie Ihr Geschlecht an.“	–
English translation of the item	An English translation of the item (note, this does not need to be a validated one)	“I get out of myself, I’m sociable.”	–	“Please indicate your gender.”	–
Questionnaire/Source	The source of the item	BFI-2	CFT-20-R	Generated ad hoc	–
Dimension the item belongs to	If the item is part of a scale, the scale the item belongs to	Extraversion	–	–	–
Response format	The format on which the response was given.	5-point rating scale	–	Single choice	–
Response labels	The labels of the response scale. If this is too long, it can be added on an additional sheet in the codebook	response_list_bfi	–	response_list_gender	–
Theoretical minimum	Lowest theoretical value the variable could take	1	55	–	–
Theoretical maximum	Highest possible value the variable could take	5	160	–	–
Coding of item	If the item is coded in the direction of the construct or not	No	–	–	–
Optional item	Indicates if the item was optional or mandatory	Yes	–	No	–
Coding of missing data	Indicates how missing data are coded	–77	NA	–	NA
Description	A brief, verbal description of the variable, especially if not an item from a survey	–	Score of the dimension	–	The time the survey was taken by the participant
Concerning study design (Examples)					
Rating source	Indicates who responded to the survey	Self	–	Self	–
Assessment wave	In studies with multiple waves, when was the item taken	First assessment	Laboratory assessment	First assessment	–
Part of survey	In larger studies, indicate block or part of survey	Personality	Cognitive ability	Demographics	–
Additional information					
Link	If available, a link to an online source of the survey/item/measure	https://search.gesis.org/instruments_tools/zis247	–	–	–
Reference	Source of the survey or assessment tool	Danner et al. (2016)	Weiβ (2019)	–	–
Comment	An additional comment in plain text, if required	–	–	–	–

Note. Information = Information that is presented in the codebook; Description = a description of that information; Individual Item = an example for an individual item from a survey; Scale Score = an example for a scale score from a questionnaire or test; Demographic = example for a demographic item; Technical = example for an item that is not responded to by the participant.

Open Data

Datendokumentation

Weiterführende Informationen sind immer sinnvoll, z.B. englische Originalformulierungen:

Response list	Coded response	Label	Translation
response_list_bfi	1	Stimme überhaupt nicht zu	Disagree strongly
	2	Stimme eher nicht zu	Disagree a little
	3	Teils, teils	Neutral
	4	Stimme eher zu	Agree a little
	5	Stimme voll und ganz zu	Agree strongly
response_list_gender	1	Männlich	Male
	2	Weiblich	Female
	3	Divers	Non-binary
	4	Keine Angabe	Prefer not to say

Note. response_list = the name of the list, which can then be referenced in each item that makes use of this scale; coded response = the numerical value in the data frame that corresponds to the “label”; translation = an English translation of the label.

Open Data

Datendokumentation

Bewertungsschema

PN-RQ (Rogge & Fincham, 2017) note	note_pnrql	<center>Fragen zu Ihnen und Ihrer Partnerschaft/Freundschaft</center>
mc_heading mc_pnrq	pre_pnrql	list_name name label
rating_button 1,7,1	pnrq_1	mc_sex 1 männlich
rating_button 1,7,1	pnrq_2	2 weiblich
rating_button 1,7,1	pnrq_3	! mc_stud 1 Psychologie
rating_button 1,7,1	pnrq_4	2 Anderes Fach: Bitte schreiben Sie welches.
rating_button 1,7,1	pnrq_5	mc_pnrq _überhaupt nicht_
rating_button 1,7,1	pnrq_6	1 _1_
rating_button 1,7,1	pnrq_7	2 _2_
rating_button 1,7,1	pnrq_8	3 _3_
submit	submit9	4 _einigermaßen_
		5 _über- wiegend_
		6 _sehr stark_
		7 _extrem stark_
note	note_pnrq2	mc_umsl 1 __trifft überhaupt nicht zu__
mc_heading mc_pnrq	pre_pnrq2	2 __trifft nicht zu__
rating_button 1,7,1	pnrq_9	3 __trifft eher nicht zu__
rating_button 1,7,1	pnrq_10	4 __trifft eher zu__
rating_button 1,7,1	pnrq_11	5 __trifft ziemlich zu__
rating_button 1,7,1	pnrq_12	6 __trifft vollkommen zu__
rating_button 1,7,1	pnrq_13	origine...
rating_button 1,7,1	pnrq_14	Schlecht
rating_button 1,7,1	pnrq_15	Langweilig
rating_button 1,7,1	pnrq_16	Leer
		Schwach
		Entmutigend
		Leblos

→ Die formr-Exceldateien decken viel vom Codebook ab (auch wenn Sie nicht unbedingt direkt selbsterklärend sind)

→ Für das Praxisprojekt reichen die formr-Exceldateien als Codebook; eine bessere Alternative für zukünftige Projekte ist die formr Datei aufzubereiten und zu ergänzen (z.B. mit dieser Vorlage: <https://osf.io/xhaey>)

Open Data

Datendokumentation

Zugang zu Präregistrierung, Open Data und reproduzierbaren Skripten	0	nicht vorhanden, oder keine persistente URL
	1	vorhanden, mit persistenter URL; am Anfang des Methodenteils (Hinweis: OSF-Links sind persistent; private Webseiten/Uniseite normal nicht). Falls intern präregistriert, bzw. Daten deponiert wurden, wurde ein entsprechender Hinweis gemacht, wo es zu finden ist (z.B. beim Dozenten).

Open Data

Grenzen: Personenbezogene Daten

Was sind "personenbezogene Daten"?

- "Data that directly can identify a person : Name, address , email address , fingerprints , date of birth , genetic data"
 - "But also unique combinations of other data that allow to identify a single person"
 - "who can be identified, directly or indirectly, by means reasonably likely to be used by [...] any [...] natural or legal person" (aus der DSGVO)
 - e.g., wer ist der männliche, 46-jährige Student im ersten Psychologie-Semester an der LMU?
-
- Wenn Teile der Daten nicht veröffentlicht werden können, heißt das nicht, dass der gesamte Datensatz nicht veröffentlicht werden kann → Legitime Datenschutzbedenken und Open Data schließen sich nicht aus!

Was sind "besonders sensible Daten"? → § 9 (1) DSGVO

- rassistische und ethnische Herkunft
- politische Meinungen
- religiöse oder weltanschauliche Überzeugungen
- Gewerkschaftszugehörigkeit,
- genetische Daten
- biometrischen Daten zur eindeutigen Identifizierung einer natürlichen Person
- Gesundheitsdaten
- Daten zum Sexualleben oder der sexuellen Orientierung → Wenn nicht nötig, nicht erheben!
Ansonsten besondere Maßnahmen notwendig.

Open Data

Grenzen: Personenbezogene Daten

Do's and Don'ts (Meyer, 2018)

- DON'T promise to destroy your data
- DON'T promise not to share data
- DON'T promise that research analyses of the collected data will be limited to certain topics
- DO get consent to retain and share data
- DO incorporate data retention and sharing clauses into IRB templates
- DO be thoughtful when considering risks of re identification
- DO consider working with a data repository
- DO be thoughtful when selecting a data repository

Open Data

Grenzen: Personenbezogene Daten

Anonymisierung

- "Anonymisieren ist das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder **nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft** einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können." (Bundesdatenschutzgesetz)
- Was "unverhältnis großer Aufwand" ist, kann von mehreren Faktoren abhängen (wer versucht es? ist es in 10 Jahren einfacher als jetzt?)
- Während es also keine 100% Garantie zur Anonymität gibt, ist das auch nicht nötig, solange man in einer Risikoabschätzung zu dem Schluss kommt, dass es ausreichend unwahrscheinlich ist (unter der Berücksichtigung der Folgen, die eine Deanoymisierung hätte)

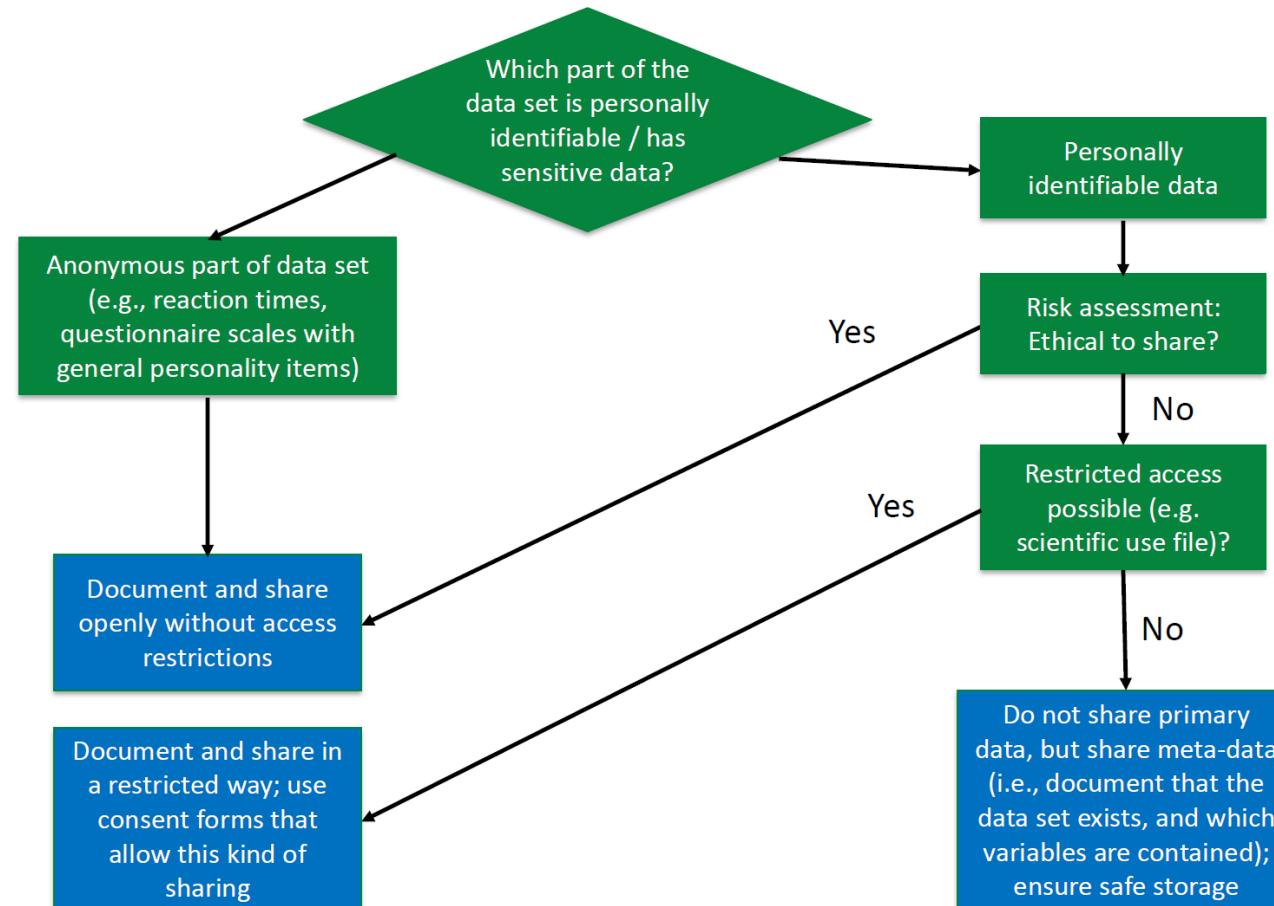
Open Data

Grenzen: Personenbezogene Daten

Risikoeinschätzung

- Wer könnte Interesse an einer Deanonymisierung haben? Wer hat zusätzliche Informationen und ein Motiv jemanden zu identifizieren?
- Vielleicht ist der CIA-Hacker mit dem Supercomputer nicht der wahrscheinlichste Angreifer und wir sollten uns viel mehr um die studentische Hilfskraft sorgen, welche vollen Zugriff auf die Daten hat und die Kommilitonen leicht identifizieren kann
- Besondere Vorsicht ist bei dyadischen Daten geboten: Romantische Partner haben zum Beispiel viele zusätzliche Informationen (z. B. Metadaten wie "Wann hat mein Partner die Fragebögen ausgefüllt?"), und wahrscheinlich ein gewisses Motiv

Open Data



Übung: Wie einfach ist eine Deanonymisierung?

Bitte Umfrage ausfüllen: <https://tellmi.psy.lmu.de/formr/anonymity>

Ihr Alter

Ihr Geschlecht

weiblich männlich anderes/keine Angabe

Bitte geben Sie Ihren Beziehungsstatus an.

Single Offene Beziehung Feste Beziehung Verlobt Verheiratet/In eingetragener Lebenspartnerschaft
 Anderer Keine Angabe

Haben Sie neben Ihrem Studium einen Nebenjob?

ja	nein
----	------

Wohnen Sie in München?

ja	nein
----	------

Haben Sie schonmal etwas anderes als Psychologie studiert?

ja	nein
----	------

Haben Sie an einer oder mehrerer Nachprüfungen teilgenommen?

ja	nein	keine Angabe
----	------	--------------

Haben Sie einen persönlichen Bezug zum Nahostkonflikt?

ja	nein	keine Angabe
----	------	--------------

Nutzen Sie Instagram?

ja	nein	keine Angabe
----	------	--------------

Nennen Sie eine peinliche Sache über sich, die Sie im Rahmen der Vorlesung preisgeben würden (kann, muss aber nicht wahrheitsgemäß sein).

Anonymisierungs-Möglichkeiten

Einzigartige Kombinationen erkennen und verhindern

Erkennen:

```
library(dplyr)

# let's try age and gender (1 = male, 2 = female)
unique_combos <- data %>%
  group_by(alter, geschlecht) %>%
  summarise(n = n()) %>%
  arrange((n))

# how many unique combinations do we have with age and gender?
table(unique_combos$n)

# show some of the critical ones
unique_combos[unique_combos$n == 1,]
```

Anonymisierungs-Möglichkeiten

Einzigartige Kombinationen erkennen und verhindern

Verhindern:

- "binning", z.B. Alter --> Alterskategorien, mehrere seltene Antworten --> in einer Antwortkategorie "andere Antworten" kombinieren → Danach überprüfen, ob das ausreichend war, d.h. keine einzigartigen Kombinationen mehr vorhanden sind!
- "fuzzing", z.B. +- 0.5 Standardabweichungen Rauschen hinzufügen (v.a. bei kontinuierlichen Antworten relevant, für Variablen die in Analysen eingehen nicht besonders geeignet)
- "deleting", z.B. Freitextantworten, Datumsangaben, Uhrzeiten, Variablen die für die Analysen nicht relevant sind

→ Diese Maßnahmen verändern und reduzieren den Wert eines Datensatzes. Unbedingt dokumentieren und eine gute Balance finden zwischen Anonymität und Nachnutzbarkeit der Daten!

Anonymisierungs-Möglichkeiten

Einzigartige Kombinationen erkennen und verhindern

Verhindern:

```
# let's make age bins
data$Alterskategorie[data$alter < 20] <- "17-19 Jahre"
data$Alterskategorie[data$alter > 19 & data$alter < 25] <- "20-24 Jahre"
data$Alterskategorie[data$alter > 24 & data$alter < 30] <- "25-29 Jahre"
data$Alterskategorie[data$alter > 29] <- "30-53 Jahre"

table(data$Alterskategorie)

## check whether we did a good job, and repeat until there are no unique entries left

## Done? Good! Save anonymized data in a separate file to share
write.csv(data, file = "raw_anonym/data_anonymized.csv", row.names = FALSE)
```

Schritt 1: Überlegen und prüfen, ob es Variablen gibt, die einzigartige Kombinationen bilden könnten und damit die Anonymität Ihrer Teilnehmer gefährden

Schritt 2: Hinreichend anonymen Datensatz erstellen

- Rohdaten in R einlesen (d.h. nicht die Rohdaten selbst verändern! Diese lokal immer unaufgetastet behalten.)
- in R ggf. einzigartige Kombinationen durch binning, fuzzing oder deleting verhindern (wenn es hierbei Hilfe in R braucht, bitte melden)
- in R Pseudonym aus den Daten löschen (`data$pseudonym <- NA`)
- Neuen anonymen Datensatz abspeichern

Schritt 3: Daten mit Codebook veröffentlichen

- Anonymen Datensatz im OSF-Projekt hochladen (wo auch die Präregistrierung liegt)
- formr Exceldateien im OSF-Projekt hochladen (wo auch die Präregistrierung liegt)