Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR ERZIEHUNGSWISSENSCHAFT

Foto: „Hamburg Überseequartier“ von A Peach / Lizenz CC BY 2.0 / Ausschnitt des Originals

02.12.2021

Workshop LEO 2018: Einführung in die Datennutzung und Sekundäranalysen

Karola Cafantaris, Gregor Dutz, Anke Grotlüschen

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR ERZIEHUNGSWISSENSCHAFT

LEO 2018:
Kompetenztest und statistischer Hintergrund

Ablauf der Befragung

- Gesamtdauer: 49 Minuten (Durchschnitt)
- Fragebogen (CAPI, ∅ 36 Minuten)
- Kompetenztest (Paper and Pencil, ∅ 13 Minuten)
- Kompetenztest umfasst vier Testhefte; es werden aber nicht alle Testhefte von den jeweiligen Befragten bearbeitet

Kompetenztest (Assessment)

- Pen & Paper
- Lese- und Schreibkompetenz (Literalität) in der deutschen Sprache

Ablauf des Kompetenztests

- Alle Befragten bearbeiten das sog. Rätselheft
- Rätselheft wird vor Ort ad hoc ausgewertet
- Beim Unterschreiten einer bestimmten Anzahl von richtigen Antworten wird eines von drei zufällig ausgewählten Testheften mit einfacheren Aufgaben zur Bearbeitung vorgelegt
- Das Umfrageinstitut wertet die Testhefte zentral nach festgelegten Regeln aus

Ziel des Kompetenztests

- **Messfehlerfreie Messung** der Lese- und Schreibfähigkeit (latente Variable) durch die Testaufgaben (manifeste Variablen).
- Genaue Beschreibung der **Bevölkerung** und von Sub-Populationen hinsichtlich der Lese- und Schreibfähigkeit.

Statistische Aufarbeitung der Testdaten

- Nutzung von **Item Response Theory (IRT)** statt klassischer Testtheorie (KTT).
- Die getesteten Personen erhalten nicht einen Kompetenzwert zugeordnet, sondern jeweils **zehn sog. plausible Werte (PVs)**.
- Diese Werte auf der LEO-Skala werden den Alpha-Levels zugeordnet

→ Bei der statistischen Auswertung der Daten sind bestimmte Verfahren einzuhalten.

Warum IRT? (Hartig und Frey 2013)

- Verortung von Itemschwierigkeit und Personenfähigkeit auf einer gemeinsamen Skala
- Multi-Matrix- bzw. Booklet-Designs sind möglich (nicht alle erhalten alle Aufgaben, adaptives Testen)
- Verlinkung von Studien auch bei einer Teilmenge von Items möglich

Warum PVs? (Lüdtke und Robitzsch 2017)

- Schätzung latenter Variablen ist messfehlerbehaftet
- Unverzerrte Schätzung der latenten Variable über Kovariate (Hintergrundmodell, im Kontext fehlender Werte: Imputationsmodell)
- Unbekannte Ausprägung der latenten Variable als spezieller Fall fehlender Werte -> „multiple Imputation“ plausibler Werte
- Die Schätzung von plausiblen Werten basiert also auf Testergebnis und Angaben im Fragebogen (Kovariate oder Hintergrundvariablen)

Beispiel

Bei identischen Antwortmustern würden Personen aus Team Grün (wahrscheinlich) höhere PVs erhalten als Personen aus Team Blau.

Voraussetzung:
Teamzugehörigkeit befindet sich im Hintergrundmodell.

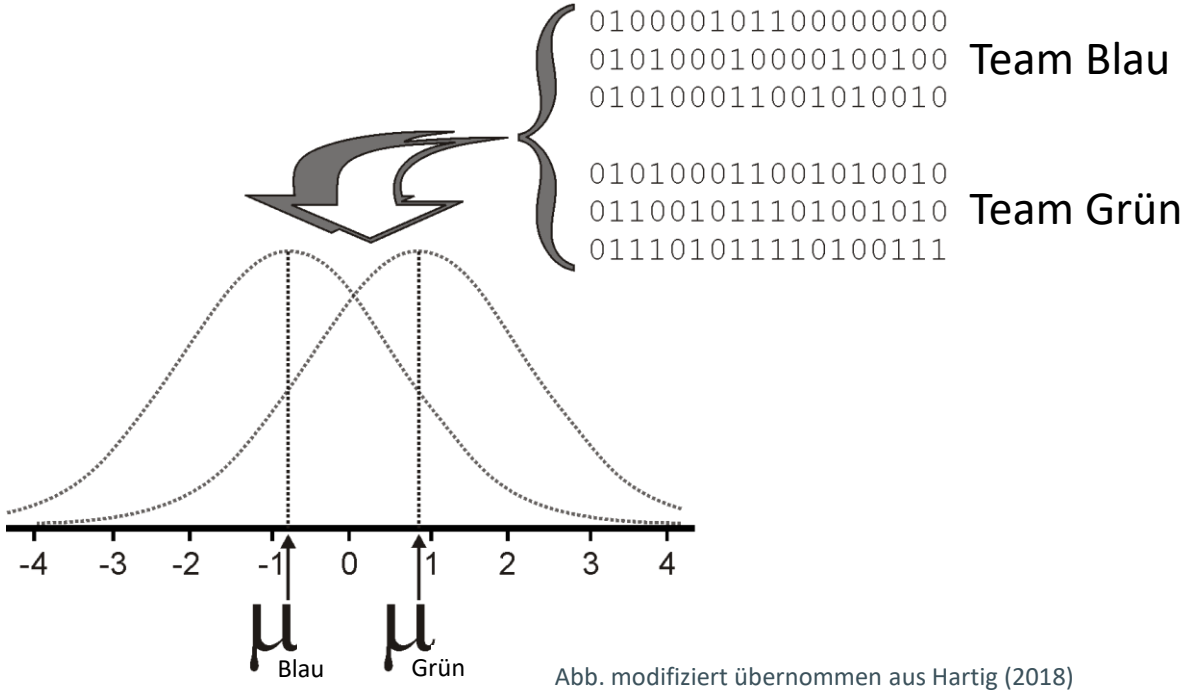

Abb. modifiziert übernommen aus Hartig (2018)

Berechnung/Ziehung von PVs

- Für jede*n Befragte*n werden in einer latenten Regression Wahrscheinlichkeitsverteilungen der wahren Fähigkeit berechnet.
- Diese Verteilungen sind abhängig von dem beobachteten Antwortmuster und den Hintergrundvariablen.
- Aus diesen Verteilungen werden zufällig die PVs gezogen.
- Je mehr PVs, desto genauer die Schätzung (in der Praxis genügen 10 PVs).

Eigenschaften von PVs

- Personen mit dem selben Antwortmuster können verschiedene PVs besitzen (wegen der zufälligen Ziehung von PVs und Unterschieden hinsichtlich der Kovariate).
- PVs sind nicht für die individuelle Diagnostik geeignet.
- PVs sind insbesondere für die genaue Beschreibung der Bevölkerung und von Subgruppen bei großen repräsentativen Stichproben geeignet (z.B. PISA, PIAAC, LEO).

Zehn plausible Werte pro Person

pv1 PV1	pv2 PV2	pv3 PV3	pv4 PV4	pv5 PV5	pv6 PV6	pv7 PV7	pv8 PV8	pv9 PV9	pv10 PV10
36.72107	37.08123	34.52205	33.14974	38.05584	38.96436	39.73280	35.41001	37.67147	36.00534
43.32304	36.37175	37.80881	39.26371	35.96375	33.59580	39.58914	40.58752	38.77825	36.57909
61.20253	60.30724	58.14160	55.12747	52.01094	54.81030	63.25451	61.67935	57.95383	63.39755
47.67577	49.22437	49.97880	52.11530	52.54056	48.18736	46.90980	49.94250	49.32013	49.70734
66.71850	57.80692	60.26352	59.26558	62.66790	57.69072	62.96557	63.10341	58.95238	68.31580

alpha_pv1 PV1 (Alpha-Level)	alpha_pv2 PV2 (Alpha-Level)	alpha_pv3 PV3 (Alpha-Level)	alpha_pv4 PV4 (Alpha-Level)	alpha_pv5 PV5 (Alpha-Level)	alpha_pv6 PV6 (Alpha-Level)	alpha_pv7 PV7 (Alpha-Level)	alpha_pv8 PV8 (Alpha-Level)	alpha_pv9 PV9 (Alpha-Level)	alpha_pv10 PV10 (Alpha-Level)
3	3	3	3	3	3	4	3	3	3
4	3	3	3	3	3	3	4	3	3
5	5	5	5	5	5	5	5	5	5
4	5	5	5	5	5	4	5	5	5
5	5	5	5	5	5	5	5	5	5

Analysen mit PVs

- Beachtung bestimmter Rechenregeln bei der Nutzung von PVs
- Entsprechend der Anzahl von PVs (hier: 10) mehrmalige Durchführung jeder statistischen Analyse mit jeweils einem PV
- Normalverteilte Punktschätzer (z.B. Mittelwerte oder Regressionskoeffizienten) als arithmetisches Mittel
- Varianzen müssen nach anderen Formeln berechnet werden

Analysen mit PVs

- Nutzung von Statistikprogrammen
  - R (mit den Paketen *survey* und *mitools*)
  - Stata (mit den Befehlen *svy* und *mi*)
  - SPSS (durch Vervielfältigung des Datensatzes)
- **Die Nutzung nur eines plausiblen Werts oder die Bildung eines Durchschnitt-PVs führt zu falschen Ergebnissen!**

Falsche Berechnungen mit PVs

Falsches Vorgehen führt zu einer Unterschätzung der Standardfehler.

	Rubin‘s Rule (richtig ☺)		ØPV (falsch ☹)	
	Wert	SE	Wert	SE
PV Mean	51,90	0,20	51,90	0,16
Intercept	41,65	0,84	41,65	0,77
Var 1	2,15	0,33	2,15	0,29
Var 2	4,08	0,25	4,08	0,23
Var 3	-1,01	0,16	-1,01	0,13

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR ERZIEHUNGSWISSENSCHAFT

LEO 2018:
Auswertung der Daten mit R

Grundlegendes zu R: Variablen

Speichern von Variablen (oder von anderen Objekten) mit dem Zuweisungoperator (<–):

```
abc <- "a"
xyz <- 123
abcxyz <- c(abc, xyz)
```

Grundlegendes zu R: Funktionen

Nutzung von Funktionen zum Durchführen bestimmter Aufgaben/Berechnungen:

```
Funktionsname(arg1 = …, arg2 = …, …)

c(abc, xyz) # kombiniert mehrere Variablen in Vektor

zahlen <- c(1, 2, 3, 4) # Vektor mit vier Zahlen
mean(x = zahlen) # Durchschnitt des Vektors zahlen
```

Grundlegendes zu R: Hilfe und Dokumentation

In R ist Hilfe mit dem ? zu erhalten:

```
?c
```

```
?mean
```

leo_doku - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | leo_doku

211118 Workshop.Rmd

Knit | Run

```
<https://www.rstudio.com/products/rstudio/download/#download>

96
97  R muss mindestens in der **Version 4.1** installiert sein,
    RStudio mindestens in der **Version 1.4**.
98
99  Sollte R oder RStudio bereits in einer älteren Version
    installiert sein, können die neueren Versionen ebenfalls
    über die Links oben heruntergeladen und einfach installiert
    werden[^2].
100
101 Haben Sie R und RStudio installiert, können Sie RStudio nun
    starten.
102
103 [^2]: Im Fall von R wird die neuere Version **zusätzlich**
    zu der älteren Version installiert, RStudio sollte jedoch
    automatisch die neuere Version nutzen.
104
105 ----------------------------------------------------------
```

137:17 | R-Pakete | R Markdown

Console | Terminal | R Markdown | Jobs

R 4.1.2 · C:/Users/Gregor/NC Arbeit/R/leo_doku/

```
----
> write_xlsx(x = m1_tab, path = "ErgebnisseReg1.xlsx")

> ## ------------------------------------------------------------
----
> m2_tab <- leo_regresult(model = m2_reg, conf.int = TRUE)

> m3_tab <- leo_regresult(model = m3_reg, conf.int = TRUE)

> ergebnisse <- list(Modell1 = m1_tab, # Der Name links vom = wird
 der Arbeitsblattname
+                    Modell2 = m2_tab,
+                    Mo .... [TRUNCATED]
```

Environment | History

Import Dataset | 421 MiB | List

R | Global Environment

Data

df	7192 obs. of 27 variables
df.sd	List of 2
ergebnisse	List of 3
f001_level_abs	List of 6
f001_level_rel	List of 6
level_abs	List of 6
level_f001_rel	List of 6
level_rel	List of 6
m1_reg	List of 10
m1_tab	11 obs. of 7 variables
m2_reg	List of 10

Files | Plots | Connections | Packages | Help | Tutorial | Viewer

Refresh Help Topic

R: Arithmetic Mean | Find in Topic

mean {base} R Documentation

Arithmetic Mean

Description

Generic function for the (trimmed) arithmetic mean.

Usage

```
mean(x, ...)

## Default S3 method:
mean(x, trim = 0, na.rm = FALSE, ...)
```

Arguments

x	An R object. Currently there are methods for numeric/logical vectors and date, date-time and time interval objects. Complex vectors are allowed for trim = 0 only.

Arithmetic Mean

Description

Generic function for the (trimmed) arithmetic mean.

Usage

```
mean(x, ...)

## Default S3 method:
mean(x, trim = 0, na.rm = FALSE, ...)
```

Arguments

`x`	An `R` object. Currently there are methods for numeric/logical vectors and date, date-time and time interval objects. Complex vectors are allowed for `trim = 0`, only.
`trim`	the fraction (0 to 0.5) of observations to be trimmed from each end of `x` before the mean is computed. Values of trim outside that range are taken as the nearest endpoint.
`na.rm`	a logical value indicating whether `NA` values should be stripped before the computation proceeds.
`...`	further arguments passed to or from other methods.

Grundlegendes zu R: Kommentare

Zeilen mit einem vorangestelltem # werden nicht ausgeführt. Kommentare dienen zur Dokumentation des Codes:

```
# Dies ist ein Kommentar

mean(x = zahlen) # Befehle vor dem # werden ausgeführt
```

Ausgangspunkt

- R (mindestens Version 4) und RStudio sind installiert
- In RStudio wurde ein neues Projekt angelegt
- Der Datensatz ist im Projektverzeichnis
- Benötigte Pakete sind installiert

Vorgehen bei der Datenauswertung in R

- Laden benötigter Pakete
- Laden des Datensatzes
- Auswahl benötigter Variablen
- Datenaufbereitung
- Erstellung eines survey-Designobjekts
- Durchführen der Analysen
- (Export der Ergebnisse)

Laden benötigter Pakete

```
library(leo)
library(tidyverse)
library(haven)
library(labelled)
library(survey)
library(mitools)
library(questionr)
library(writexl)
```

Laden des Datensatzes

```
datenpfad <- "ZA6266_v1-0-0.sav"

df <- read_sav(file = datenpfad)
```

Auswahl benötigter Variablen

```
df <- df |> select(pgewges, # Gewichtungsvariable
    contains("pv"), # wählt alle PV-Variablen aus
    f001, altgr5, # Gender, Alter
    schulab, zuf001) # Schulab., Zufriedenheit
```

Hinweis: Sollen im Verlauf der Analysen weitere Variablen hinzugenommen werden, müssen diese in der obigen Syntax ergänzt werden. Anschließend die Syntax von Beginn neu durchlaufen lassen.

Blick in ausgewählte Variablen

```
describe(x = df$f001)
```

```
## [7192 obs.] Geschlecht
## labelled double: 2 2 2 1 2 1 1 2 2 1 ...
## min: 1 - max: 2 - NAs: 0 (0%) - 2 unique values
## 2 value labels: [1] männlich [2] weiblich
##
##                    n     %
## [1] männlich    3129  43.5
## [2] weiblich    4063  56.5
## Total           7192 100.0
```

Datenaufbereitung

- Umwandeln in Faktorvariablen
- Recodieren und fehlende Werte setzen
- Neue Referenzkategorie setzen

Umwandeln in Faktorvariable

Hintergrund: R erkennt kategoriale Variablen nicht automatisch. Diese müssen in sog. Faktorvariablen umgewandelt werden.

```
df$f001 <- to_factor(x = df$f001)
```

Dies muss für alle kategoriale Variablen durchgeführt werden.

Hinweis: Zuerst müssen ggf. andere Veränderungen an der Variablen durchgeführt werden (bspw. Recodierungen). Erst danach erfolgt die Umwandlung in eine Faktorvariable.

Recodieren und fehlende Werte setzen

Nicht interessierende Ausprägungen können auf fehlend (NA) gesetzt werden:

```
# `keine Angabe` auf fehlend setzen
df$schulab[df$schulab == 9] <- NA
# Umwandeln in Faktor und leere Kategorien entfernen
df$schulab <- to_factor(x = df$schulab,
                        drop_unused_labels = TRUE)
```

Recodieren und fehlende Werte setzen

Bei kontinuierlichen Variablen ist das setzen fehlender Werte besonders wichtig:

```
# `keine Angabe` auf fehlend setzen
df$zuf001[df$zuf001 == 99] <- NA
```

Neue Referenzkategorien setzen

Für Regressionen können die Referenzkategorien neu bestimmt werden:

```
# Zuerst in Faktor umwandeln
df$altgr5 <- to_factor(x = df$altgr5)
# Dann die Referenzkategorie setzen
df$altgr5 <- fct_relevel(df$altgr5, "55-64 Jahre")
```

Erstellung eines survey-Designobjekts

- Definition der PVs als multiple Imputationen: Umwandlung der wide-Version (jeder PV liegt als separate Variable vor) in eine Liste mit zehn Datensätzen und jeweils einem PV.
- Definition als Survey-Daten (Gewichtung)

Erledigt das leo-Paket (fast) automatisch:

```
df.sd <- leo_svydesign(data = df)
```

Verschiedene plausible Werte im Datensatz

- `pv`: Literalität auf der kontinuierlichen LEO-Skala
- `alpha_pv`: Literalität in Alpha-Levels
- `lowlit`: Dichotomisierte Literalität (gering literalisiert/nicht gering literalisiert)

Durchführen der Analysen

Die Durchführung der Analysen mit plausiblen Werten erfolgt in zwei Schritten:

- Zunächst wird jede Analyse entsprechend der Anzahl der PVs durchgeführt (das survey-Designobjekt enthält insgesamt zehn Datensätze)
- Anschließend werden die Ergebnisse gepoolt.

Durchführen der Analysen

- Durchschnitt
- Absolute Häufigkeitstabellen
- Relative Häufigkeitstabellen
- Kreuztabellen mit relativen Häufigkeiten
- Regressionen

Durchschnitt

```
# Berechnungen werden mit automatisch 10x durchgeführt
pv_durchschnitt_temp <- with(df.sd, svymean(~pv, na.rm
  = TRUE))
# Ergebnisse werden gepoolt
pv_durchschnitt <- MIcombine(pv_durchschnitt_temp)
# Durchschnitt anzeigen
summary(pv_durchschnitt)
```

Absolute Häufigkeitstabellen

```
level_abs <- MIcombine(with(df.sd, svytotal(~alpha_pv,
  na.rm = TRUE)))
level_abs
```

```
## Multiple imputation results:
##       with(df.sd, svytotal(~alpha_pv, na.rm = TRUE))
##       MIcombine.default(with(df.sd, svytotal(~alpha_pv, na.rm = TRUE)))
##                results       se
## alpha_pva1   45.8599 12.34056
## alpha_pva2  241.4815 29.44115
## alpha_pva3  579.5603 41.51163
## alpha_pva4 1472.6011 62.87005
## alpha_pva5 4852.3392 79.31979
```

Relative Häufigkeitstabellen

```
level_rel <- MIcombine(with(df.sd, svymean(~alpha_pv,
  na.rm = TRUE)))
level_rel
```

Kreuztabellen mit relativen Häufigkeiten

```
level_f001_rel <- MIcombine(with(df.sd, svyby(~alpha_pv,
  ~f001, svymean, na.rm = TRUE)))
level_f001_rel
```

Hinweis: Die Reihenfolge der beiden Variablen `alpha_pv` und `f001` ist für die Interpretation der Ergebnisse wichtig!

```
## Multiple imputation results:
##       with(df.sd, svyby(~alpha_pv, ~f001, svymean))
##       MIcombine.default(with(df.sd, svyby(~alpha_pv, ~f001, svymean)))
##                              results          se
## männlich:alpha_pva1 0.010536909 0.003246495
## weiblich:alpha_pva1 0.002097731 0.001122735
## männlich:alpha_pva2 0.042019291 0.005878750
## weiblich:alpha_pva2 0.024894178 0.004757049
## männlich:alpha_pva3 0.086164766 0.008572246
## weiblich:alpha_pva3 0.074847681 0.007747214
## männlich:alpha_pva4 0.221013156 0.013322701
## weiblich:alpha_pva4 0.188043087 0.009730671
## männlich:alpha_pva5 0.640265878 0.013260382
## weiblich:alpha_pva5 0.710117323 0.011848377
```

Geschlecht nach Alpha-Levels

Beispiel: 71% der Frauen besitzen Literalität auf Level 5.

Kreuztabellen mit relativen Häufigkeiten

```
level_f001_rel <- MIcombine(with(df.sd, svyby(~f001,
  ~alpha_pv, svymean, na.rm = TRUE)))
level_f001_rel
```

Hinweis: Die Reihenfolge der beiden Variablen wurde hier umgedreht .

```
## Multiple imputation results:
##       with(df.sd, svyby(~f001, ~alpha_pv, svymean))
##       MIcombine.default(with(df.sd, svyby(~f001, ~alpha_pv, svymean)))
##                        results         se
## a1:f001männlich 0.8378257 0.08504799
## a2:f001männlich 0.6352832 0.05091381
## a3:f001männlich 0.5420240 0.03789877
## a4:f001männlich 0.5471509 0.02134593
## a5:f001männlich 0.4811523 0.01026848
## a1:f001weiblich 0.1621743 0.08504799
## a2:f001weiblich 0.3647168 0.05091381
## a3:f001weiblich 0.4579760 0.03789877
## a4:f001weiblich 0.4528491 0.02134593
## a5:f001weiblich 0.5188477 0.01026848
```

Alpha-Levels nach Geschlecht

Beispiel: 54% derjenigen auf Alpha-Level 3 sind männlich, 46% sind weiblich.

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR ERZIEHUNGSWISSENSCHAFT

LEO 2018:
Workshop Gruppenphase

Gruppenphase

- Ziel: Replikation von Tabellen aus dem Ergebnisband
- Jeweils Gruppen zu den vier Grundbildungsdomänen (Finanzen, Digitale, Gesundheit, Politik)
- Berechnung von relativen Häufigkeiten und Kreuztabellen mit jeweils fünf Variablen und einer Literalitätsvariablen.
- Empfehlung: Benutzen Sie `lowlit` als Literalitätsvariable, das erleichtert die Auswertung der Ergebnistabellen.

Tabelle 2: Im Alltag mitbekommen, was in Deutschland und der Welt passiert, nach Alpha-Levels (Anteile in Prozent, gewichtet) (Quelle: Universität Hamburg, LEO 2018 – Leben mit geringer Literalität.)

	α 1	α 2	α 3	α 4	über α 4	Ges.-Bev.	α 1–3	α 4 u. höher	Sig.
täglich	*27,4*	66,6	75,1	81,9	84,9	82,5	70,2	84,2	sig.
mindestens einmal pro Woche	*44,6*	*16,2*	16,1	12,8	12,4	13,1	17,6	12,5	n.s.
seltener als einmal pro Woche	*14,7*	*7,5*	*5,1*	3,3	2,0	2,8	*6,3*	2,3	sig.
seltener als einmal im Monat	*11,4*	*7,1*	*1,8*	*0,6*	*0,4*	0,9	*3,8*	*0,4*	sig.
nie	*1,8*	*2,6*	*1,2*	*0,8*	*0,2*	*0,5*	*1,6*	*0,3*	sig.
keine Angabe	*0,0*	*0,0*	*0,7*	*0,5*	*0,2*	*0,3*	*0,5*	*0,3*	n.s.
Summe	*100,0*	100,0	100,0	100,0	100,0	100,0	100,0	100,0	

Basis: Deutsch sprechende Erwachsene (18–64 Jahre), n = 7.192; kursiv und grau gesetzte Ziffern basieren auf Fallzahlen < 60; Gesamtzusammenhang zwischen Alpha 1–3 und darüber ist signifikant (Chi²-Test). Abweichung der Summen von 100 % aufgrund von Rundungseffekten.

Quelle:
Dutz/Grotlüschen (2019, S. 296)

Gruppenphase

Für jede der fünf Variablen aus den vier Domänen:

- Relative Verteilung der Antworthäufigkeiten für die Gesamtbevölkerung
- Kreuztabellen: Alpha-Levels nach Antwortkategorie

Gruppenphase

Bereitgestellte Materialien/Hilfsmittel:

- Codebook und Fragebogen
- Präsentation und PDF zur Einführung in R
- Dokument mit den Fragen/Variablen für jede Gruppe
- Musterlösung mit Variablennamen und Syntax

Gruppenphase

Gruppenphase bis **14:30 Uhr**, danach Treffen im Plenum zur Diskussion der Ergebnisse.

Sollten Sie schneller fertig sein, können Sie in der Gruppe:

- Andere Fragen aus Domänen ausprobieren
- Statt der Literalitätsvariablen das Alter, Geschlecht, etc. nutzen
- Regressionen ausprobieren (siehe PDF)
- Ideen für eigene Forschungsfragen diskutieren