4.2.2
Mittelwertbasierter Reliabilitätstest10 Manual.aux Manual.aux

Intercoderreliabilität und Expertenvalidität bei Inhaltsanalysen

Erläuterungen zur Berechnung des Reliabilitätskoeffizienten Lotus mit SPSS

Benjamin Fretwurst

18. März 2013

Reliabilitätskoeffizient Lotus (λ)[1]

Die Qualitätsdimensionen einer inhaltsanalytischen Messung werden als *Reliabilität* und *Validität* bezeichnet. Im Folgenden wird beschrieben, wie ein intuitiv verständlicher Reliabilitätskoeffizient aussehen und recht einfach mit dem Statistikpaket SPSS berechnet werden kann. Der Reliabilitätskoeffizient ist unabhängig von der Anzahl der Kodierer und kann für nominale, ordinale und metrische Variablen ausgegeben werden. Anhand der Übereinstimmung aller Kodierer mit einem «Goldstandard» wird instrumentelle Reliabilität bzw. Expertenvalidität erfasst. Der Koeffizient ist intuitiver, leichter berechenbar und direkter vergleichbar als die gängigen Alternativen, wie Krippendorfs α oder Cohens κ. Manche Fehler der Alternativen werden vermieden: Es werden nur Hauptübereinstimmungen gewertet und geringere Reliabilität bei mehr Ausprägungen, werden nicht herausgerechnet. Im Vortrag soll der Koeffizient vorgestellt und mit den gebräuchlichen Reliabilitätskoeffizienten verglichen werden. Anschliessend wird die Verwendung ein SPSS-Makros zur Berechnung des Koeffizienten vorgestellt.

[1] Da der vorgeschlagene Koeffizient noch keinen Namen hat, muss ein Henkel dran: *Lotus*. Als Symbol schlage ich λ vor. So lässt es sich leichter über den Koeffizienten sprechen.

Inhaltsverzeichnis

1 Messtheoretische Grundlagen des Reliabilitätstests

1.1 Reliabilität und Validität

Reliabilität wird in der Regel mit «Zuverlässigkeit» gleichgesetzt. Reliabilität beschreibt rechnerisch die Ähnlichkeit gemessener Werte bei wiederholter Messung des gleichen Untersuchungsmaterials. Trotzdem Reliabilität zu den methodischen Standardkonstrukten gehört, ist nicht ganz klar, was sie misst und messen soll. Wird die Qualität der Kodierer, Eigenheiten des Relimaterials, die Qualität eines Codbuchs oder die der Schulung gemessen? Soll der Koeffizient die Kodiererschulung steuern, angeben wie stark die Kodierung durch das Codebuch determiniert ist oder (rituell) die Datenqualität einer Studie vermitteln?

Als Qualitätsmerkmal eines Messinstruments gibt die Reliabilität an, wie viel der Gesamtstreuung einer Stichprobenvariablen nicht auf Messfehlerstreuung, sondern auf die wahre Streuung der Variablen in der Grundgesamtheit zurückgeführt werden kann (vgl. Bühner 2011). Wenn also ein Messinstrument keine Messfehlerstreuung erzeugt, dann ist der Reliabilitätskoeffizient gleich eins, liegt demnach bei 100%. Wäre die gesamte Varianz einer Stichprobenvariablen allein auf Messfehlerstreuung zurückzuführen, wäre die Reliabilität gleich null. In die Messfehlerstreuung geht die Qualität der Kodierer, des Codebuchs und der Kodiererschulung ein. In dieser Form gibt der Reliabilitätskoeffizient valide Auskunft über den Anteil der Messfehlerstreuung in den Daten – kann aber keine klare Auskunft über die Qualität des Codebuchs selbst geben. Dazu muss (und kann (nur) mit Lotus) der Messfehleranteil der Kodierer herausgerechnet werden. Eine Reliabilitätsmessung vor einer Schulung ist nicht sinnvoll. Welchen Anteil die Kodiererschulung an der Reliabilität hat, kann nicht eindeutig bestimmt werden. Es ist daher eine Frage der Ehre, dass Entscheidungen und Anweisungen aus der Kodiererschulung im Codebuch transparent gemacht werden.

Mit dem Begriff «Stichprobenvariable» soll unterstrichen werden, dass die Reliabilität eines Messinstruments nur als Messfehlerstreuung einer gemessenen Variable bestimmt werden kann. Zu der Messfehlerstreuung kommt im ungünstigen Fall ein systematischer Bias hinzu, der nicht durch den Reliabilitätskoeffizienten abgebildet wird. Durch eine solche Verzerrung sind die Schätzer nicht mehr erwartungstreu. Weicht zum Beispiel der Mittelwert einer Variablen systematisch vom zu schätzenden Parameter in der Grundgesamtheit ab, kann dieser Bias eines Messinstruments nicht anhand der Stichprobe bestimmt werden. Dieser Aspekt fehlerhafter Messung wird nicht durch das Kriterium der Reliabilität beschrieben, sondern durch das Konzept der Validität. Eine naheliegende Möglichkeit die Validität einer Inhaltsanalyse zu bestimmen, ist die Expertenvalidität. Dazu werden die Kodierentscheidungen der Kodierer mit denen eines oder mehrerer Experten verglichen. Das Relimaterial, das die Kodierer für den Relitest (mancher würde sagen: während des Relitests) verkodieren, wird ebenfalls kodiert: 1. vom Forschungsleiter, 2. von einem Expertenteam oder 3. in gemeinsamer Abstimmung mit Experten und Kodierern. Die erste Variante ist die einfachste, bietet aber am wenigsten Kontrolle. Externe Experten das Ralimaterial verkodieren lassen, bietet ein deutlich stärkeres Aussenkriterium. Allerdings wird es nicht immer leicht sein, Experten zu finden, die das Codebuch gründlich lesen und zum Teil umfangreiche Relikodierungen durchführen. Eine gute Variante ist die Bildung eines kleinen Expertenteams, das über die Kodierung des Relimaterials deliberativ entscheidet.

1.2 Gehalt

Als Menschen mit stochastischem Weltbild (es gibt Zufall) halten wir es für möglich, dasss Kodierungen zufällig richtig geraten und wollen diese «Zufallstreffer» herausrechnen. Das kommt uns spätestens nach einer Statistikausbildung normal vor, ist es aber nicht. Kodierer würfeln nicht. Sicher gibt es Fehler und merkwürdige oder einfach schlampige Entscheidungen, aber die gehorchen nicht sicher den Gesetzen des Zufalls. Was hier kontrolliert werden soll, ist eine Qualitätsdimension, die über Gültigkeit und Zuverlässigkeit hinaus geht: Es geht um Gehalt. Damit ist nicht die monetäre Entschädigung von Kodierern gemeint, sondern der Informationsgehalt von Messungen. Dass das eine zusätzliche Qualitätsdimension ist, soll folgendes Beispiel

illustrieren: Einer mehr oder weniger interessanten Forschungsfrage folgend soll «Prominenz» erhoben werden. Das ist nie ganz einfach, aber es wird mit der Unterscheidung in A-Promis, B-Promis, C- und D-Promis begonnen. Während der Kodiererschulung stellt sich heraus, dass das zu ambitioniert war und diese Kodierung nicht in dieser Feinheit reliabel. Die persönliche Wertung der Kodierer lässt auch keine besonders valide Unterscheidung von B- und C-Promis zu. Daher wird nur noch zwischen internationaler und nationaler Prominenz unterschieden. Das ist zuverlässiger und damit auch valider. Die Gültigkeit dieser Unterscheidung liegt klarer auf der Hand. Was wir verloren haben ist Informationsgehalt. Die Steigerung der Reliabilität und Validität wird mit der Reduktion der des Inforamtionsgehalts erkauft.

Dieses Prinzip steht im Hintergrund, wenn wir die mögliche Anzahl zu kodierender Ausprägungen aus der Reliabilität «herausrechnen» wollen. Hat eine Variable wenig Kategorien, ist sie viel leichter reliabel kodierbar als die Operationalisierung des gleichen Konstruktes mit einer Variable mit vielen Kategorien. Das kommt uns unfair vor, bedeutet aber nur, dass wir den Verlust des Informationsgehaltes kompensieren wollen. Technisch und mathematisch schlägt sich diese Überlegung in sämtlichen Standardisierungsalgorithmen wieder. Begründet werden diese zwar in der Regel mit dem Herausrechnen zufällig richtiger Kodierungen, aber Kodierer würfeln nicht und zufällig richtig ist auch richtig. Es geht um Gehalt. Für Lotus gibt es auch standardisierte Varianten. Dabei wird auch der «Zufall» herausgerechnet. Mit diesem gängigen Verfahren entspricht auch LOTUS dem zaghaften Versuch der Vergleichbarkeit des Gehalts entgegen zu kommen. Eine bessere Idee, den Gehalt von Variablen messbar zu machen steht noch aus, aber ich überlege die ganze Zeit.

1.3 Eigenschaften

Die Reliabilitätskoeffizienten nach Lotus sind in der Regel höher als Krippendorfs α oder Cohens κ. Das liegt vor allem daran, dass Lotus nicht aus der Kombination sämtlicher Vergleiche bestimmt wird, sondern als Übereinstimmung mit Vorgabewerten. Lotus gibt exakt den Anteil der Übereinstimmungen mit dem am häufigsten kodierten Wert an und Lotus-Goldstandard (LGS) den Anteil der Übereinstimmungen mit dem Goldstandard.

Für die Tabelle **??** wurden Variablen mit vorgegebenen Eigenschaften erstellt. Dabei wurde per Zufallsvorgabe die Anzahl von Übereinstimmungen unter den Kodierern vorgegeben und der Anteil der Übereinstimmungen mit dem Goldstandard. Für diese Montecarlo Studie wurden 100 Kodiereinheiten und 10 Kodierer angelegt. Die Tabelle zeigt einen Ausschnitt der Variablen und resultierenden Lotus-Koeffizienten. Anhand des Variablennamens kann abgelesen werden, wie die Variable konstruiert wurde. Nach dem Buchstaben i folgt die Vorgabe für die Interkoderreliabilität nach Lotus in Prozent. Hinter gs steht der Prozentsatz der Übereinstimmungen mit einem zusätzlichen Vorgabewert, der als der «wahre» Wert betrachtet werden soll. An der letzten Stelle ist angegeben, wie viele Kategorien die Variable hat.

Wenn man sich die Variablen mit gleichem i ansieht, stellt man für i4 gleiche Werte für Lotus fest. Das muss auch so sein, weil der Goldstandard nicht beachtet wird. Bei i5 bis i9 sind die Werte leicht unterschiedlich, weil sie eben zufällig erzeugt wurden. Die Zeile i5gs8c6 ergibt einen Lotus-Wert von .67 also eine Übereinstimmung von 67%. Das kommt daher, dass 50% der Werte per inhaltlicher Vorgabe gleich sind, aber noch zufällige Übereinstimmungen hinzu kommen. Diese Werte sind in S-Lotus herausgerechnet. Der Erwartungswert von S-Lotus liegt genau bei der Vorgabe.[2] Der LGS-Wert ergibt sich aus dem Produkt der Interkoderreliabilität und der GS-Vorgabe. In S-LGS kann dieser Zusammenhang klarer wiedererkannt werden. Bei i8gs8c6 stimmt es trotz Zufallsprozess genau: S-Lotus ist .8 und S-LGS ist .64. Für diese Version des Manuals habe ich realistische Werte für die Fallzahlen und Coder genommen. Mit sehr grossen Stichproben von etwa 10000 Kodiereinheiten und 100 Kodierern stimmen die Werte genau. Ich überlege noch, was ich für die Eigenschaften von Lotus präsentieren will.

[2] S-Lotus ist nur dann zu hoch, wenn der Goldstandard sehr hoch ist und die Interkodervorgabe sehr klein (<4).

Reliabilität und Validität mit Lotus

Variablen	Lotus*	S-Lotus**	LGS***	S-LGS**
i1gs9c6	,37	,24	,26	,11
i2gs9c6	,39	,27	,31	,17
i3gs9c6	,43	,31	,38	,25
i4gs8c6	,52	,42	,41	,30
i4gs9c6	,52	,42	,48	,38
i5gs8c6	,54	,45	,49	,39
i5gs9c6	,61	,53	,59	,50
i6gs8c6	,67	,60	,59	,50
i6gs9c6	,64	,57	,58	,49
i7gs8c6	,76	,72	,68	,61
i7gs9c6	,78	,73	,72	,66
i8gs8c6	,84	,80	,70	,64
i8gs9c6	,83	,80	,77	,72
i9gs8c6	,92	,90	,74	,69
i9gs9c6	,91	,90	,85	,82

* Anteil der Übereinstimmungen
** Standardisierte Koeffizienten sind zufallsbereinigt
*** Anteil der Übereinstimmungen mit dem Goldstandard

2 Der Koeffizient Lotus

Die Reliabilität, also die Zuverlässigkeit der inhaltsanalytischen Messung, wird hier als «Intercoderreliabilität» verstanden. Sie wird als prozentuale Übereinstimmung aller Kodierer mit dem am häufigsten kodierten Wert pro Fall bestimmt[3]. Diese Rechenweise weicht von den Angaben paarweiser Vergleiche (z. B. nach Holsti und Derivaten:[4] wie Scotts Pi, Cohens Kappa oder Krippendorffs α) ab (vgl. Kolb 2004).

Der am häufigsten kodierte Wert steht für den vom Instrument implizierten Wert und soll als «Hauptübereinstimmung» bezeichnet werden. Durch Lotus wird nur die Hauptübereinstimmung gemessen. Wenn Kodierer gemeinsam andere Werte kodiert haben, geht das in Lotus nicht ein. Koeffizienten, die auf Paarvergleiche aufbauen, haben zwangsläufig die Schwäche, dass Gleichkodierungen von falschen Werten ebenfalls positiv in die Reliabilität eingehen. Die Frage nach der Übereinstimmung mit dem am häufigsten kodierten Wert ist Reliabilität im ganz wörtlichen Sinne, denn eine von allen Kodierern gleichermaßen falsche Kodierung ist zuverlässig falsch.

Durch den vorgeschlagenen Koeffizienten wird nicht zwischen Variablen mit wenigen gegenüber vielen Kategorien unterschieden. Die erhöhte «Schwierigkeit» (vgl. Wirth 2001) schlägt sich also in schlechteren Reliabilitätswerten nieder. Variablen mit höherem Schwierigkeitsgrad haben faktisch eine geringere Reliabilität. Das darf nicht «herausgerechnet» werden, wie es bei z.B. Fleiss κ und Krippendorffs α geschieht (vgl. Krippendorff 2004). Wenn durch mehr Ausprägungen die Wahrscheinlichkeit von Fehlern steigt, dann ist die Reliabilität tatsächlich geringer. Die Idee hinter der Relativierung an der Wahrscheinlichkeit ist eine andere: hier soll der Gehalt einer Variable mit beachtet werden. Mehr Abstufungen einer inhaltsanalytischen Variable sind gehaltvoller als die reliablere Reduzierung bis zur Dummyvariablen. Der Gehalt einer differenziert abgestuften Variable ist aber eine nur qualitativ bestimmbare Eigenheit und kann nicht standardisiert gegengerechnet werden. Wahrscheinlichkeitserwägungen helfen dabei nicht weiter, weil auch Kodierer nicht würfeln. Es liegt also nicht einmal abstrakt betrachtet ein Wahrscheinlichkeitsexperiment vor. Der Forschungsleiter muss den gewünschten Gehalt der Variablen festlegen und darauf hin arbeiten, dass dieser Gehalt auch reliabel gemessen wird.

Vergleiche zwischen nominalen Variablen sind einfach, weil die kodierten Werte gleich sind oder eben nicht. Schwieriger scheint der Vergleich, wenn nicht bloss identische, sondern auch ähnliche Kodierungen als zuverlässig betrachtet werden sollen. Krippendorffs α enthält in der Routine für ordinale und metrische Variablen Distanzmasse, an denen die Reliabilität relativiert wird (vgl. Krippendorff 2004). Der hier vorgeschlagene Koeffizient Lotus stellt Vergleiche bei gegebener *Toleranz* an. Damit kann und muss bei ordinalen und metrischen Variablen angegeben werden, um wie viele Einheiten die Kodierungen maximal abweichen dürfen, bis sie nicht mehr als zuverlässig gelten. Wird beispielsweise die Dauer eines Nachrichtenbeitrags in Sekunden erfasst, könnten 3 Sekunden als tolerierbar gelten. Reliabilitätswerte für metrische Variablen müssen daher immer mit ihrer Toleranz ausgewiesen werden. Damit ist der hier vorgeschlagene Koeffizient auch bei metrischen Variablen deutlich transparenter als die Krippendorffsche quadrierte Distanzenarithmetik.

3 Übereinstimmungen mit einer Goldstandard-Vorgabe als Expertenvalidität

Die Richtigkeit oder Gültigkeit einer Messung wird durch die Validität der Messung angegeben. Als Indikator für Validität kann in Anlehnung an den Reliabilitätskoeffizienten die prozentuale Übereinstimmung aller Kodierer mit einer Goldstandard-Vorgabe herangezogen werden. Folgt man Werner Früh (2007) entspricht die Übereinstimmung der Kodierer mit dem Forschungslei-

[3] Dieses Prinzip hat auch Steffen Kolb 2004 vorgeschlagen, allerdings nur für nominale Variablen.

[4] Bei mehreren Kodierern ergeben sich durch die Vielzahl der Kombinationsmöglichkeiten Redundanzen, die in den verschiedenen Reliabilitätskoeffizienten durch komplizierte, Umrechnungen herausgerechnet werden müssen.

ter eine Form der Expertenvalidität. Dabei wird das Relitestmaterial auch vom Forschungsleiter kodiert, der das kodiert, was gemessen werden soll. Im Unterschied zur Übereinstimmung mit dem Forschungsleiter wird ein Goldstandard als Vorgabe betrachtet, die unterschiedlich zustande kommen kann. Dabei ist die Forschungsleitervorgabe die einfachste Variante. Der Forschungsleiter nimmt dabei die Rolle eines Experten ein, der erkennt was eigentlich zu kodieren wäre. Ein Goldstandard als valide Vorgabe kann zusätzlich zum Forschungsleiter durch zusätzliche Experten, in reflektierter Kodierarbeit zusammen mit erfahrenen Kodierern oder durch eine ergänzende Befragung von Rezpienten vorgenommen werden, um festzustellen, wie ein zu kodierender Inhalt interpretiert wird, und demnach auch von den Kodierern zu interpretieren ist. Wenn keine Experteneinschätzungen oder Befragungen im Budget liegen, sollte der Forschungsleiter mindestens selbst das Untersuchungsmaterial kodieren, um an einem Minimal-Goldstandard prüfen zu können.

Die Intercoderreliabilität ist maximal so gross wie die Übereinstimmung mit dem Goldstandard. Wenn der Goldstandard immer den gleichen Wert vergeben hat wie die Mehrheit der Kodierer, dann sind die Intercoderreliabilität und Expertenvalidität gleich. Ist im Goldstandard ein anderer Wert vergeben als von der Kodierer-Mehrheit, liegt die Intercoderreliabilität immer über der Übereinstimmung mit dem Goldstandard.[5] Hier wird das logische Verhältnis zwischen Reliabilität und Validität mathematisch wiedergegeben: Reliabilität ist die *notwendige*, aber nicht *hinreichende* Voraussetzung für Validität. Eine unzuverlässige Messung kann nicht gültig sein. Zuverlässig ungültige Messungen sind hingegen möglich.

[5] Nur in dem ganz speziellen Fall, dass alle Kodierer eine Variable für eine Kodiereinheit unterschiedlich kodiert haben, es aber zwischen einem Kodierer und dem Forschungsleiter eine Übereinstimmung gibt, kann die Intercoderreliabilität unter der Übereinstimmung mit dem Forschungsleiter liegen.

4 Umsetzung mit SPSS

Im Unterschied zu Krippendorffs α kann der hier vorgestellte Reliabilitätskoeffizient vergleichsweise einfach mit SPSS bestimmt werden. Dafür stehen SPSS-Makros zur Verfügung, die im Folgenden vorgestellt werden. Rahmenvorgabe für die Entwicklung von Lotus war: 1. keine zusätzliche Installation irgend welcher Programme und keine Umstrukturierung des Datensatzes!

Die Verwendung der «ReliMacros» ermöglicht es *Intercoderreliabilität* respektive *Expertenvalidität* für jede Variable eines Reliabilitätsdatensatzes zu bestimmen. Zusätzlich können für jeden Kodierer die Reliabilitätswerte pro Variable ausgegeben und aus Lotus herausgerechnet werden. Schliesslich kann die Zuverlässigkeit jedes Kodierers für alle Variablen bestimmt werden.

4.1 Vorbereitung der Daten

Zum Aufbau des Reliabilitätsdatensatzes: Die Kodierungen jedes Kodierers werden während der Relikodierung in jeweils einer Datendatei erfasst, die für jeden Kodierer gleich strukturiert ist. Die Variable, die die Kodiereinheit (z.B. Artikel oder Nachrichtenbeitrag) enthält, muss den Variablennamen ‹cu› tragen (Coding Unit).[6] Die Variable, in der die Kodierernummer eingetragen wird muss «coder» heissen. Die Vorgabekodierung durch den Forschungsleiter oder eine Expertengruppe, in der Variable `coder` die Nummer 0 haben. Die Variablen im Datensatz dürfen keine Leerzeichen und keine Umlaute oder sonstige Sonderzeichen ausser den Unterstrich _ enthalten.[7]

Die Dateien des Relitests sollten in einem Ordner liegen; dieser Pfad wird SPSS mit folgendem Befehl mitgeteilt (er muss also für den eigenen Relitest angepasst werden).

```
cd 'C:\Dropbox\meins\Relitest'.
```

In diesem Ordner müssen auch die Vorlage-Datei `Relitest.sps` und die Makrodatei `LOTUS.sps` gespeichert sein .

Die Dateien aller Kodierer müssen in einer Datei zusammengefasst sein.

```
add file  /file =  /file ='ReliCoder2.sav' /in = FileCoder2.
var lab FileCoder2 'ReliCoder2.sav vom 13.11.2011'.
save out RelitestGesamt.sav.
```

Die Reliabilitäten können nur für numerische Variablen durchgeführt werden. Stringvariablen müssen also bereinigt und in numerische Variablen umkodiert werden.[8]

4.2 Verwendung des Makros

Bevor die Reliabilitätstest ausgeführt werden können, muss (immer wenn SPSS neu aufgerufen wurde) die Makrodatei gestartet werden. Wenn die Datei `LOTUS.sps` im durch `cd ...` festgelegten Verzeichnis liegt, genügt folgender Befehl:

```
Insert file ='LOTUS.sps' syntax = interactive.
```

Soll das Makro an einer anderen Stelle gespeichert werden, kann und muss der entsprechende Speicherpfad angegeben werden.

Es gibt grundsätzlich zwei Varianten für die Vergleiche. Zum einen den Reliabilitätstest für Variablen mit Kategorienvorgaben und zum anderen die mittelwertbasierten Tests.

---

[6] In der Dialogfeldvariante kann angegeben werden, welche Variable die Kodiereinheit und welche die Kodiererkennung enthält. Soll das Makro über den Syntax benutzt werden, sollten diese Variablen vorher schlicht kopiert und als CU und CODER bezeichnet werden. Das noch in das Makro zu integrieren wäre übertrieben gewesen.

[7] SPSS erlaubt in der Standardnutzung inzwischen Variablen mit Leerzeichen und Umlauten, aber die Makrosprache von SPSS kann damit nicht umgehen.

[8] Dazu eignet sich der Befehl `autorecode` von SPSS, der über das Menü «Transformieren» erreichbar ist (Automatisch umkodieren:.

Die Bezeichnung für dieser Testvarianten ist etwas merkwürdig. Anfangs wurden dieser Test für nominale Variablen gebaut. Aber auch ordinale Variablen und metrische mit nur wenigen Ausprägungen sollten mit Hilfe dieser Version getestet werden. Nur wenn bei im Prinzip gleichen Kodierentscheidungen andere Werte eingegeben werden können, (weil es keine exakten Vorgaben gibt) ist die zweite Variante angebracht, die weiter unten erklärt wird. Zum Beispiel sind Variablen für Nachrichtenfaktoren häufig mindestens ordinal: geringer Schaden, mittlerer Schaden, hoher Schaden. Selbst klassisch metrische Variablen, wie die Anzahl von Bildern pro Zeitungsartikel hat erwartbar wenige Ausprägungen und wird mit der Standardvariante für Kategorienvorgaben bestimmt.

Im Unterschied dazu könnte man die Länge eines Nachrichtenbeitrags in Sekunden messen. Dabei ist eine Abweichung von ein paar Sekunden zu verschmerzen. Eine Orientierung an der am häufigsten kodierten Sekundenzahl wäre nicht sinnvoll. Eine solche Variable wird dann mit der mittelwertbasierten Variante bestimmt, bei der nicht die Übereinstimmung mit der am häufigsten kodierten Sekundenzahl als Maß genommen wird, sondern die Übereinstimmung mit dem Mittelwert bei gegebener Toleranz von ein paar Sekunden.

4.2.1 Reliabilitätstest mit exakten Kategorienvergleichen

Die meisten Variablen aus Inhaltsanalysen haben zwei oder ein paar vorgegebene Kategorien. Varialben können einzeln oder als Liste verschiedener Variablen vorgegeben werden. Dazu gibt man nach dem Schlüsselwort `!LOTUS` mit `varlist =` Jede Variable an, die auf die gleiche Weise getestet werden soll. Also:

```
!Lotus varlist = v1 v2 v5 v7 v23.
```

Da die Eingabe jeder Variablen etwas mühsam ist, gibt es die Möglichkeit eine Reihe Variablen durch ihren Begin und ihr Ende anzugeben. Also:

```
!Lotus beg = v1 end = v23.
```

Beide Varianten sind kombinierbar, wobei am Ende der `varlist` ein Schrägstrich stehen muss, damit SPSS weiss, dass die Liste vollständig ist. Also zum Beispiel:

```
!Lotus varlist = v1 v4 /beg = v7 end = v17.
```

4.2.2 Mittelwertbasierter Reliabilitätstest

Die Reliabilität von kontinuierlichen Variablen ist nur dann sinnvoll berechen- und angebbar, wenn Toleranzen für abweichende Kodierungen angegeben werden. Damit sind vor allem metrische Variablen mit kontinuierlichen Werten gemeint, bei denen Toleranzen für Übereinstimmungen angegeben werden müssen.[9] Soll zum Beispiel der Umfang von zu kodierenden Seiten in cm^2 angegeben werden, ist es nicht sinnvoll, jede Abweichung als unzuverlässige Kodierung einzustufen. Mit dem Schalter `level = M` und `TOL =` *Wert* wird die *tolerierbare* Abweichung der Kodierer von der *durchschnittlichen* Kodierung als *reliable* Kodierung betrachtet. Daher muss in `tol` angegeben werden, mit welcher Toleranz Abweichungen vom Mittelwert noch als richtige Kodierung betrachtet werden. Im Beispiel wird bei Variable `var32` jede Kodierung als richtig betrachtet werden, die nur um zwei Einheiten vom Mittelwert der Kodierungen abweicht. Alle Variablen, die mit der gleichen Toleranz getestet werden sollen, können wiederum als Einzeleingaben über zum Beispiel `varlist = Artikelumfang Bildumfang Kastenumfang /level = M tol = 2` getestet werden.

```
!Lotus  varlist = var32/level = M tol = 2.
```

[9] Metrische Variablen mit nur sehr wenig Ausprägungen und ohne Messtoleranz sollten mit dem Test für exakte Kategorienvergleiche getestet werden.

4.3 Standardisierte Koeffizienten

Neben den einfachen Übereinstimmungen unter den Kodierern, also Lotus und den Übereinstimmungen mit dem Goldstandard (LGS), werden durch das Makro standardisierte Werte ausgegeben (zur Diskussion siehe **??**). Die standardisierten Koeffizienten geben den Anteil der nichtzufällig richtigen Kodierungen wieder. S-Lotus wird wie folgt berechnet (und S-LGS entsprechend):

$$\text{S-Lotus} = \frac{Lotus - \frac{1}{Kategorien}}{1 - \frac{1}{Kategorien}}$$

4.4 Schalter für Sonderfälle

Das Lotus-Makro kennt einige Schalter. Mit ihrer Hilfe wird es ermöglicht auf typische Sonderfälle einer Inhaltsanalyse einzugehen. Die ersten Schalter wurden schon bei den mittelwertbasierten Tests vorgestellt: `level` und `tol`. Im Folgenden soll auf die Sonderfälle und Schalter eingegangen werden.

4.4.1 Toleranzen bei Problemen an Kategoriengrenzen

Bei Mittelwertbasierten Tests wird auf jeden Fall die Angabe von Toleranzen benötigt, da eine exakte Übereinstimmung eher unwahrscheinlich und daher nicht zu verlangen ist. Aber auch bei Variablen mit wenigen vorgegebenen Kategorien kann es sinnvoll sein Toleranzen anzugeben. Das ist dann der Fall, wenn Kategorien an den Rändern unscharf sind. Wenn es bei (mindestens ordinalen) Variablen häufig Fälle gibt, die sich in der Nähe von Kategoriengrenzen bewegen, hilft eine Rekodierung der Variablen oft nicht weiter, weil so eine Reduktion der Kategorien nur die Anzahl der Grenzen verringert, aber nicht das Grenzproblem angeht. Das geht dann auch noch erheblich auf Kosten des Gehalts. Mit der Angabe von Toleranzen kann das Grenzproblem besser gelöst werden. Dabei wird bei jeder Kodierung im Vergleich zur häufigsten Kodierung und zum Goldstandard einfach die Nachbarkategorie noch als richtig betrachtet. Daher kann auch für den Standardfall exakter Kategorienvergleiche mit `tol = 1` dem Grenzproblem begegnet werden (level = C braucht nicht angegeben werden, weil es standard ist).[10]

4.4.2 Missings und Filter

In der Regel können fehlende Angaben wie normale Kodierungen betrachtet werden. Es werden daher auch in der Standardeinstellung von Lotus alle fehlenden Werte als gleichwertige Kodierungen betrachtet. Wenn ein Kodierer entgegen der Mehrheit eine Eigenschaft nicht kodiert hat, ist das ein Fehler der so in Lotus eingeht. Dieses Verhalten kann abgeschaltet werden indem der Schalter `miss = 1` gesetzt wird. Dann werden Fälle mit fehlenden Werten nicht in die Berechnung von Lotus mit einbezogen.

Anders verhält es sich mit bedingten Kodierungen bzw. Filtern. Sollen zum Beispiel Akteure und ihre Eigenschaften verkodiert werden, kann es vorkommen, dass es in einigen Kodiereinheiten keine Akteure gibt und daher die dazugehörigen Eigenschaften nicht sinnvoll kodiert werden können. Das kann zu unter- und überschätzten Relikoeffizienten führen. Ist bei einem Artikel eindeutig kein Akteur vorhanden, fehlen alle Werte seine Folgevariablen, und das ist richtig. Allerdings wollen wir nicht das Eindeutige Fehlen in vielen Eigenschaftskategorien messen, sondern die Zuverlässige Beschreibung von Akteuren. Daher gibt es bei Lotus einen Schalter `NoCode`. Aufgrund eines Filters fehlende Werte in bedingten Kategorien sollten auf einen Wert festgesetzt werden, zum Beispiel -7. Also zum Beispiel:

[10] Die so bestimmte Reliabilität gilt nicht für die Ausgangvariable, weil die Grenzprobleme ja tatsächlich bestehen. Die Reliabilität würde also als zu hoch angegeben werden. Die Variable darf also genau genommen nur als zusammengefasste Variable in die Auswertungen eingehen. Alternativ besteht die Möglichkeit die ursprüngliche Reliabilität und die mit Toleranz anzugeben und zu diskutieren. Grenzprobleme reduzieren die Reliabilitätswerte selbst dann, wenn die Fehler zu vernachlässigen sind. Diskussionsfreudige und mutige Wissenschaftler verwenden trotz der Reliabilitätsangabe mit Toleranz die Ausgangsvariable und sind sich der etwas zu hohen Reliabilitätsangabe bewusst.

```
!Lotus varlist = v6a v6b v6c v6d v6e /NoCode = -7.
```

Es ist nicht ratsam die Fehlenden Werte aufgrund von bedingten Kodierungen als system fehlende Werte stehen zu lassen, da richtige bedingte Auslassung von Variablen von vergessenen Variablen nicht mehr unterschieden werden könnten. Also sollte vor dem Relitest die Filterung in den bedingten Variablen in einen NoCode umgerechnet werden. Also zum Beispiel:

```
do if (v6 = 0).
recode v6a to v6e (sysmiss = -7).
end if.
!Lotus varlist = v6a v6b v6c v6d v6e /NoCode = -7.
```

4.4.3 Hierarchische Variablen

Gelegentlich kommt es vor, dass Ausprägungen von Variablen hierarchisch geordnet sind. das ist typisch bei Themen oder Akteuren. Solche hierarchischen Ausprägungen sollten nach dem Hotelzimmermuster kodiert werden, also die erste Stelle für die übergeordnete Kategorie, die zweite für die erste Unterkategorie und die dritte für die feinste Differenzierung. Zum Beispiel könnten die Hunderter einer Themenkodierung für die gröbste Themenunterteilung vorgehalten werden und die die Zehner für Unterthemen: 100 Politik, 200 Wirtschaft, 300 Gesellschaft ... 101 Parteipolitik, 102, Sachpolitik :., 201 Finanzwirtschaft, 202 Arbeitsplätze usw.

Im Makro ist für solche Fälle der Schalter `Hiera` vorgesehen. Wird zum Beispiel `Hiera = 2` angegeben, wird die Reliabilität für Hunderterebene ausgegeben, also nur, ob Kodierer zuverlässig zwischen Politik, Wirtschaft und Gesellschaft unterscheiden können (schwer genug). Die Verrechnung der beiden Koeffizienten ist bis auf Weiteres in Lotus nicht vorgesehen, da das eher inhaltliche Entscheidungen sind. Eine Möglichkeit besteht darin, den Durchschnitt aus dem Reliabilitätskoeffizienten für die feinste Ebene mit dem oder den Werten für höhere Ebenen zu berechnen.

4.4.4 Seltene Phenomene

Seltene Phänomene (rare phenomena) sind so rar, dass sie in Reliabilitätstests selten gut getestet werden können, da dazu das Relimaterial viele Kodiereinheiten umfassen müsste.[11] Wenn viel Relimaterial vorliegt, können Rare-Phenomena-Probleme genauer untersucht werden. Im Grunde sind es für den Reliabilitätstest eher Rare-Phenomena-Glücksfälle, da sie zu hohen Reliabilitätswerten führen. Zum Beispiel kommt der eine oder andere auf die Idee den Nachrichtenwert «Sexualität» in Nachrichten nachzugehen (vgl. Fretwurst 2008). Sex kommt aber in Nachrichten sehr selten vor. Kodierer erkennen recht sicher in 98 Prozent der Fäll liegt kein Sex vor. Das ist natürlich auch Zuverlässigkeit, aber ich möchte herausfinden, wie zuverlässig Kodierer Sex und Erotik erkennen, wenn sie denn mal vorkommt. Dafür gibt es den Schalter `RaPh`. Wenn der für eine oder mehrere Variablen auf 1 gestellt ist, werden nur solche Kodiereinheiten in die Berechnung von Lotus einbezogen, die nicht bei allen Kodierern eine 0 haben. Damit wird der Frage nach der zuverlässigen Kodierung seltener Phänomene Rechnung getragen. Fälle in denen nur ein Kodierer Sex zu erkennen vermeint, gehen in die Berechnung mit ein.

4.4.5 Kategorienzahl vorgeben

Für die standardisierten Koeffizienten muss die Anzahl der möglichen Kategorien bekannt sein. Im Makro kann nur die Anzahl der verwendeten Kategorien pro Variable bestimmt werden. Kommen in einem Reliablitätstest nicht alle Kategorien vor, kann die Anzahl der Kategorien vorgegeben werden.

[11] Neuerdings werden aber viele Kodierungen für die Programmierung von automatischen Inhaltsanalysen vorgenommen, deren Übereinstimmung mit den automatischen Kodierungen dann auch getestet werden soll.

4.4.6 Tabelle für alle Variablen

Bei jedem Durchlauf des Makros werden für alle angegebenen Variablen Tabellen für die Interkoderreliabilität und den Goldstandard, sowie die standardisierten Koeffizienten ausgegeben. In der letzten Zeile stehen immer auch die Gesamtwerte. Diese Werte sind nicht gedacht um die «Gesamtreliabilität» einer Inhaltsanalyse anzugeben, sondern um Kodiererleistungen miteinander vergleichen zu und ein Feedback geben zu können. Nachdem das Makro für alle interessierenden Variablen durchgelaufen ist, kann eine Gesamttabelle ausgegeben werden. Dabei werden die Reliwerte nicht neu berechnet. Das ist wichtig, da sonst die Wirkung der unterschiedlichen Schalter für die unterschiedlichen Variablen wieder aufgehoben würde. Wenn zum Beispiel zwischen der ersten Variable `Zeitung` und der letzten `Emotion Freude` keine Stringvariable besteht, kann man mit folgendem Befehl eine Tabelle für alle Variablen ausgeben:[12]

```
!LotusTable beg = Zeitung end = EmoFreu.
```

Mit diesem Verfahren sind auch Analysen für Variablenblöcke möglich und sinnvoll. So könnte man für Formalia und Akteure jeweils eine Gesamtreliabilität ausgeben, um in Veröffentlichungen mit wenig Platz für ausführliche Relidiskussionen wenigstens Durchschnitte für Variablengruppen angeben zu können.

Das noch eher experimentelle Makro `!LotusAggr` liefert eine kompakte Zusammenfassung der Reliabilitätstest mehrerer Variablen ohne die Angabe für alle Kodierer. Dieses Makro funktioniert wie das der Tabellen, krankt aber an Umsetzungproblemen mit SPSS.

[12] Dieser Befehl kenn nur den Schalter `gs`, der auf 0 gesetzt werden muss, wenn es leider keinen Goldstandard gibt.

5 Lotus per Dialogfeld

Lotus kann auch über das Menü gesteuert werden, wobei alle Funktionen verwendet werden können. Dazu muss das Dialogfeld Lotus.spd installiert werden.[13] Nach der Installation ist Lotus über das Menü Analysieren im Unterpunkt «Deskriptive Statistik» zu finden.

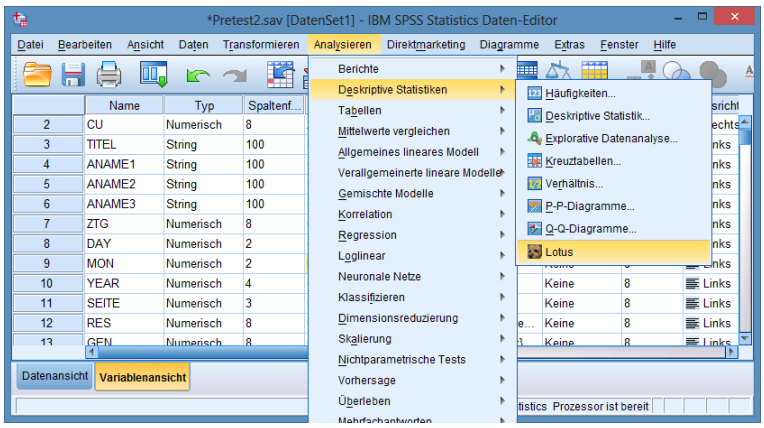

Abbildung 1: Lotus im Menü

Die Verwendung entspricht dem des Makros. In der Quellliste «Variablen» sind nur Variablen aufgeführt, die keine Stringvariablen sind. Dadurch ist sichergestellt, dass keine Stringvariablen in die Zielliste mit aufgenommen werden. Das Menü muss jeweils für Variablen mit unterschiedlichen Eigenschaften separat ausgeführt werden. Mit jedem Durchlauf werden Variablen für die Reliwerte in der Datendatei abgelegt. Damit kann am Ende eine Tabelle für alle interessierenden Variablen ausgegeben werden. Für diese abschließende Analyse muss das Kästchen «Nur Tabellen» angeklickt werden.

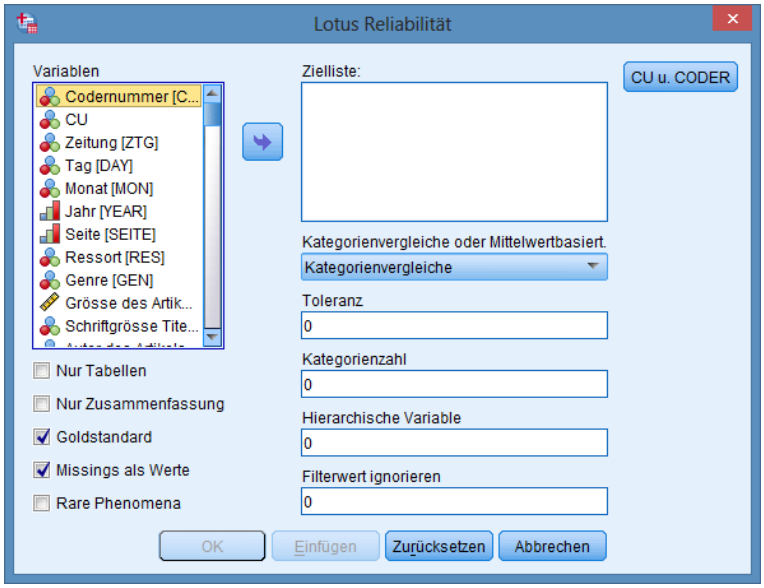

Abbildung 2: Lotus über das Dialogfeld

[13] Bis zur SPSS-Version 20 kann das Dialogfeld nur mit Administratorenrechten installiert werden. Ab Version 21 lässt speichert SPSS das benutzerdefinierte Dialogfelder in einem Benutzerordner ab.

Literatur

Bühner, Markus (2011): *Einführung in die Test- und Fragebogenkonstruktion.* Pearson.

Fretwurst, Benjamin (2005): *Notwendige Angaben zu Auswahlverfahren.* Theoretische Überlegungen und eine empirische Auswertung der Dokumentationspraxis in der Kommunikationswissenschaft. Mit: Gehrau, Volker/Weber, René In: Gehrau, Volker/ Fretwurst, *Benjamin*/Krause, Birgit/Daschmann, Gregor (2005, Hrsg.): *Auswahlverfahren in der Kommunikationswissenschaft.* Köln, von Halem-Verlg. S. 32–51.

Fleiss, Joseph L. (1981): The measurement of interrater agreement. In: ders., *Statistical methods for rates and proportions.* John Wiley & Sons. 212–236.

Früh, Werner (2007): *Inhaltsanalyse.* UTB.

Fretwurst, Benjamin (2008): *Nachrichten im Interesse der Zuschauer.* Eine konzeptionelle und empirische Neubestimmung der Nachrichtenwerttheorie. UVK-Verlag. Konstanz.

Kolb, Steffen (2004): Verlässlichkeit von Inhaltsanalysedaten. In: *Medien- & Kommunikationswissenschaft.* 52 Jg. 2004/3.

Krippendorff, Klaus (2004). Reliability in content analysis: Some common misconceptions and recommendations. In: *Human Communication Research*. 30, 411–433.

Krippendorff, Klaus (2011): *Computing Krippendorff's Alpha-Reliability.* `http://www.asc.upenn.edu/usr/krippendorff/mwebreliability4.pdf`.