Kapitel 11

Statistische Auswertungsansätze für die empirische Unterrichts-forschung

……

11.3 Zum Umgang mit fehlenden Daten

Fehlende Daten sind in empirischen Untersuchungen normal, auch in der Unterrichtsforschung sind sie kaum zu vermeiden. Für den Umgang mit fehlenden Daten werden in der Praxis oft relativ einfache, aber ungünstige Verfahren eingesetzt (Lüdtke, Robitzsch, Trautwein & Köller, 2007), obwohl mit *Multiple Imputation* eine flexible Methode zur Verfügung steht (Little & Rubin, 2002; Van Buuren, 2012).

11.3.1 Einfache Verfahren

Die einfachen Verfahren umfassen u.a. die listenweise und die paarweise Löschung von Werten sowie die Mittelwert- und die Regressionswertzuschreibung. Abbildung 11.1 veranschaulicht die Prinzipien der Verfahren an einem einfachen Datensatz mit sieben Fällen (s_1, s_2, …, s_7) und drei Variablen (Y_1, Y_2, Y_3) mit fehlenden Werten (*NA*). Während die listenweise und die paarweise Löschung zu einer Reduzierung der Fälle führt (s_2 und s_5 bzw. s_5 werden gelöscht), bleibt bei Mittelwert- und Regressionswertzuschreibung die Fallzahl komplett.

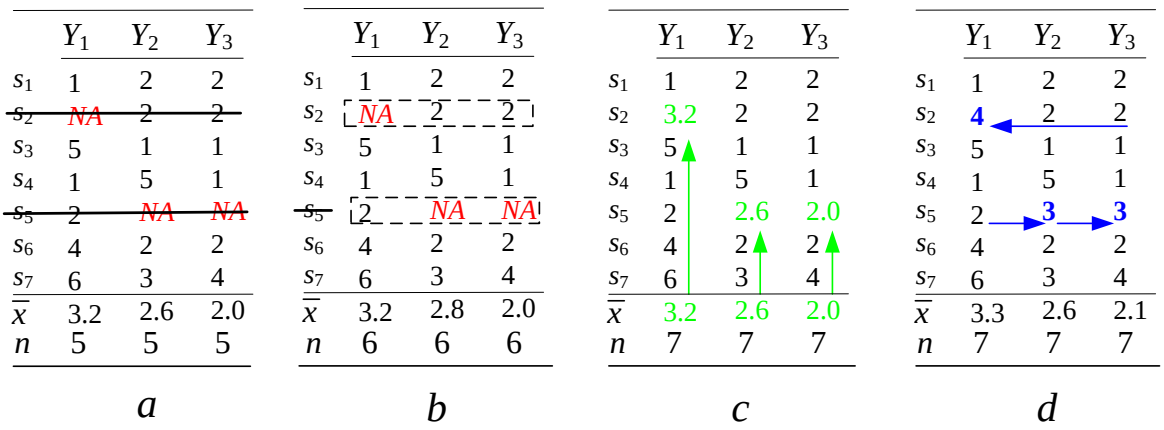


Abbildung 11.1 Prinzipien der listenweise (*a*) und paarweisen Löschung (*b*), Mittelwert- (*c*) und Regressionswertzuschreibung (*d*)

Listenweise Löschung

Die listenweise Löschung *(listwise deletion*) ist das voreingestellte Verfahren in den meisten Statistikpaketen (z. B. SPSS). Bei diesem Verfahren werden alle Fälle mit einem oder mehreren fehlenden Werten in Bezug zu abhängigen Variablen eliminiert (vgl. Abbildung 11.1a). Der Vorteil des Verfahrens ist seine Einfachheit, der Nachteil der verschwenderische Umgang mit Daten.

Paarweise Löschung

Die paarweise Löschung (*pairwise deletion*) versucht die Nachteile der listenweisen Löschung zu kompensieren. Bei diesem Verfahren werden Mittelwerte und Kovarianzen für beobachtete Daten berechnet. Die Berechnung der Mittelwerte erfolgt über die vorhandenen Daten je Variable (vgl. Abbildung 11.1b), die Berechnung der Kovarianzen erfolgt über die vorhandenen Datenpaare je zweier Variablen.

Mittelwertzuschreibung

Beim Verfahren der Mittelwertzuschreibung (*mean imputation*) werden fehlende Werte durch den Mittelwert ersetzt, getrennt nach Variablen (vgl. Abbildung 11.1c). Das Verfahren ist unkompliziert, allerdings wird die Varianz der Variablen unterschätzt, zudem zerstört es Beziehungen zwischen Variablen.

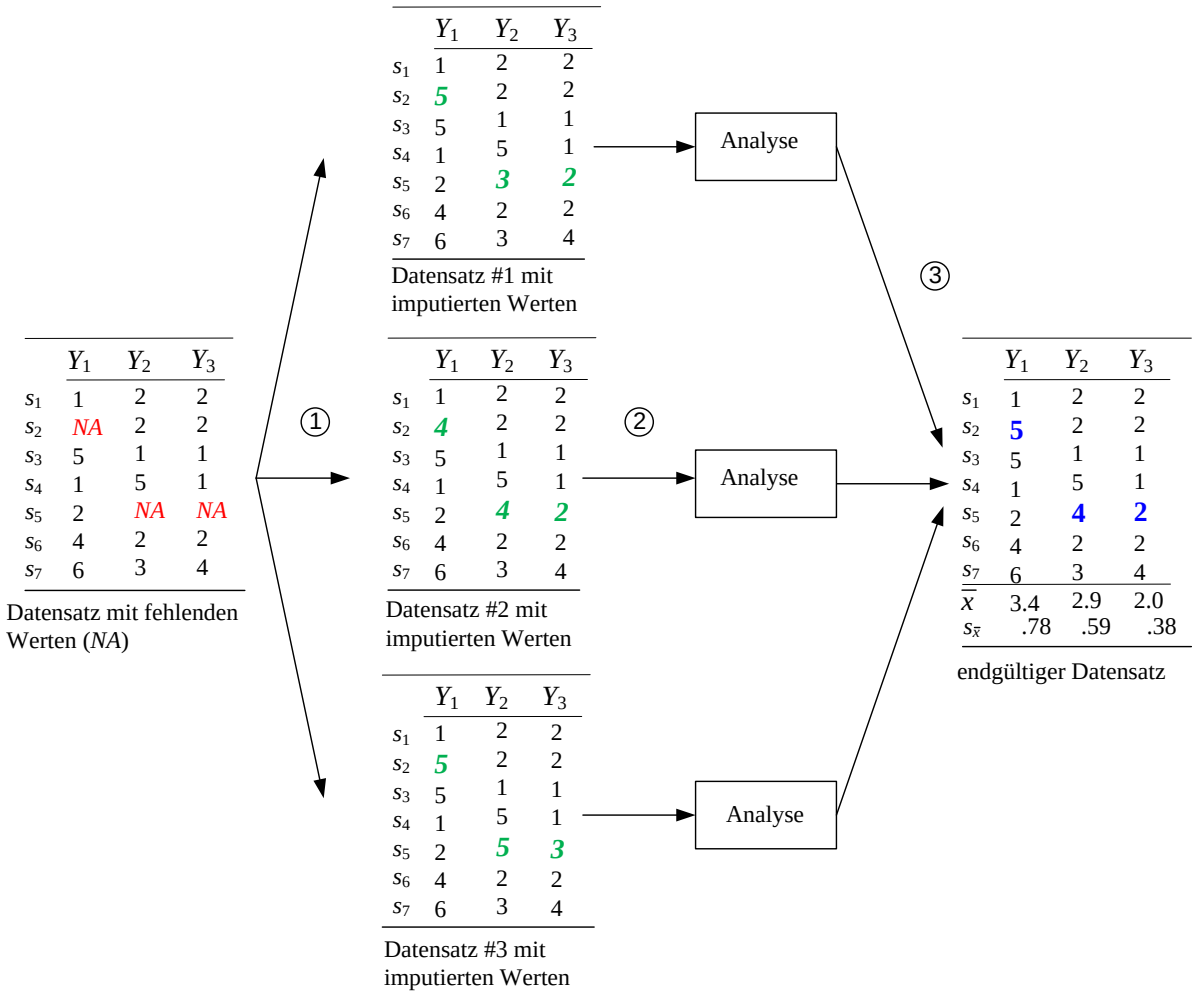
Regressionswertzuschreibung

Beim Verfahren der Regressionswertzuschreibung (*regression imputation*) werden fehlende Werte in einer (abhängigen) Variable durch berechnete Werte mit Hilfe von Regressionsrechnung ersetzt (vgl. Abbildung 11.1d), wobei ein Modell zwischen abhängiger Variable und einer oder mehreren unabhängigen Variablen zugrunde gelegt sein muss.

11.3.2 *Multiple Imputation*

Das Verfahren der multiplen Imputation geht auf Rubin (1976) zurück. Das Verfahren erzeugt $m > 1$ komplette Datensätze, damit Unsicherheiten beim Ersetzen fehlender Werte berücksichtigt werden können. Abbildung 11.2 veranschaulicht das prinzipielle Vorgehen: (1) Aus einem Datensatz mit fehlenden Werten werden z. B. $m = 3$

Datensätze mit imputierten (zugeschriebenen) Werten erzeugt. Die erzeugten Datensätze sind identisch für die empirisch beobachteten Werte; sie unterscheiden sich in den imputierten Werten. (2) Die erzeugten Datensätze werden statistischen Analysen unterzogen und (3) zu einem endgültigen Datensatz gepoolt.



	Y_1	Y_2	Y_3
s_1	1	2	2
s_2	*NA*	2	2
s_3	5	1	1
s_4	1	5	1
s_5	2	*NA*	*NA*
s_6	4	2	2
s_7	6	3	4

Datensatz mit fehlenden Werten (*NA*)

	Y_1	Y_2	Y_3
s_1	1	2	2
s_2	***5***	2	2
s_3	5	1	1
s_4	1	5	1
s_5	2	***3***	***2***
s_6	4	2	2
s_7	6	3	4

Datensatz #1 mit imputierten Werten

	Y_1	Y_2	Y_3
s_1	1	2	2
s_2	***4***	2	2
s_3	5	1	1
s_4	1	5	1
s_5	2	***4***	***2***
s_6	4	2	2
s_7	6	3	4

Datensatz #2 mit imputierten Werten

	Y_1	Y_2	Y_3
s_1	1	2	2
s_2	***5***	2	2
s_3	5	1	1
s_4	1	5	1
s_5	2	***5***	***3***
s_6	4	2	2
s_7	6	3	4

Datensatz #3 mit imputierten Werten

	Y_1	Y_2	Y_3
s_1	1	2	2
s_2	**5**	2	2
s_3	5	1	1
s_4	1	5	1
s_5	2	**4**	**2**
s_6	4	2	2
s_7	6	3	4
$\bar{x}$	3.4	2.9	2.0
$s_{\bar{x}}$	.78	.59	.38

endgültiger Datensatz

Abbildung 11.2 Verfahren der multiplen Imputation

Multiple Imputation löst zum einen das Problem fehlender Daten. Zum anderen liefert es die Lösung, damit komplette Datensätze (Datensätze mit voller Fallzahl) in die weitere statistische Verarbeitung eingehen können. Zudem löst es das Problem von zu großen und zu kleinen Standardfehlern, die bei listenweiser bzw. paarweiser Löschung, Mittelwert- und Regressionswertzuschreibung auftreten und zu Verfälschungen in nachgeschalteten statistischen Analysen (z.B. Varianzanalysen) führen können.

Das Verfahren der multiplen Imputation steht in SPSS seit Version 17 zur Verfügung. Sehr flexibel einsetzbar ist das Paket *mice* unter R (Van Buuren & Groothuis-Oudshoorn, 2011; Van Buuren, 2012; Kabacoff, 2015), das eine Vielzahl von Funktionen bereitstellt, welche

- die Inspektion von Mustern fehlender Daten,
- die Erzeugung von Datensätzen mit imputierten Werten,
- die Diagnose imputierter Werte,
- die Analyse kompletter Datensätze,
- das Pooling wiederholter Analysen,
- das Speichern und Exportieren von Datensätzen mit imputierten Werten,
- die Generierung von Datensätzen,
- den Einbau benutzerspezifischer Imputationsmethoden

erlauben.

Bereitgestellte Vignetten im *mice-Paket* unter R führen das Verfahren der multiplen Imputation Schritt für Schritt ein, mit Vignetten für einfache Datensätze bis hin zu Datensätzen aus komplexen (multilevel) Versuchsplänen.

11.x Literatur

Kabacoff, R. L. (2015). *R in action*. Shelter Island, NY: Manning.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York. Wiley.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. Probleme und Lösungen. *Psychologische Rundschau, 58*(2), 103–117.

Van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, Fl: CRC Press.

Van Buuren, S., Groothuis-Oudshoorn, K. (2011). `mice`: Multivariate **i**mputation by **c**hained equations in `R`. *Journal of Statistical Software*, *45*(3), 1–67.