

Evaluation Randomisiert-Kontrollierter Studien und Experimente mit R

Auswertung Multipel Imputierter Daten: Die Rubin-Regeln

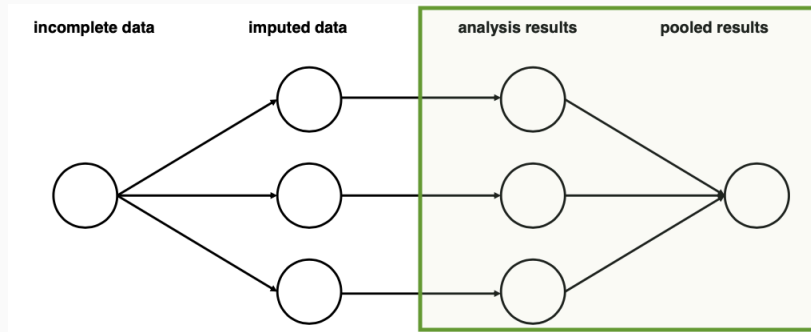
Prof. Dr. David Ebert & Mathias Harrer

Graduiertenseminar TUM-FGZ

Psychology & Digital Mental Health Care, Technische Universität München

Die 3 Variationsquellen

Wie komme ich mit m Imputationssets zu einem gemeinsamen Analyseergebnis?



adaptiert von Van Buuren & Groothuis-Oudshoorn (2011).

→ **Lösung:** Durchführung der selben Analyse in allen m imputierten Datensets parallel, dann werden alle m relevanten Parameter/Schätzwerte zu einem Wert aggregiert (“**gepoolt**”).

Dabei müssen wir miteinbeziehen, dass unsere Schätzwerte **mit Unsicherheit behaftet** sind. Diese Unsicherheit hat bei MI mindestens zwei Gründe:

1. Die Teilnehmenden der Studie stellen nur eine Stichprobe der untersuchten Studienpopulation dar, sind also mit Stichprobenfehler (“**sampling error**”) behaftet.
2. Die Daten enthalten **fehlende Werte**, deren Schätzung **unsicher** ist. Diese Unsicherheit wird dadurch reflektiert, dass multipel imputierte Werte sich zwischen Imputationssets unterscheiden (können).

Es sei Q ein zu schätzender wahrer Wert (oder ein Vektor von Werten) der Population (z.B. Populationsmittelwert, Regressionskoeffizienten, ...).

Aufgrund der zuvor genannten Gründe ist Q unbekannt und muss durch einen Schätzer \hat{Q} angenähert werden. Dies ist nur durch die beobachteten Werte Y_{obs} möglich.

Der Erwartungswert (d.h. die bestmögliche Annäherung) von Q gegeben Y_{obs} ist (Van Buuren, 2018, Kapitel 2.3.2):

$$E(Q|Y_{\text{obs}}) = E\left(E(Q|Y_{\text{obs}}, Y_{\text{mis}}) \mid Y_{\text{obs}}\right)$$

d.h. der Durchschnittswert der (imputierten) Schätzungen des Mittelwerts von Q über alle multiplen Imputationen hinweg.

Kombination von Punktschätzungen:

Es sei \hat{Q}_ℓ die Schätzung von Q im ℓ -ten von m Imputationssets. Der “gepoolte” Schätzwert von Q ist damit:

$$\bar{Q} = \frac{1}{m} \sum_{\ell=1}^m \hat{Q}_\ell$$

Aggregation von Punktschätzungen bei MI

Um Punktschätzungen (z.B. Parameter der Wahrscheinlichkeitsverteilung wie Mittelwert & Standardabweichung; Regressionsgewichte, etc.) zu poolen, wird in **jedem Imputationsset** der Wert des Punktschätzers **berechnet**, und daraufhin der **Mittelwert über alle Imputationsets** gebildet.

Aber wie kann die **Unsicherheit** (Varianz) von Q in MI-Daten geschätzt werden? Die Varianz von Q gegeben Y_{obs} besteht aus **zwei Komponenten**:

$$V(Q|Y_{\text{obs}}) = \underbrace{E\left(V(Q|Y_{\text{obs}}, Y_{\text{mis}}) \mid Y_{\text{obs}}\right)}_{\substack{\text{Mittelwert d. Varianzen über alle MI-Sets} \\ \rightarrow \text{Within-Variance } (\bar{U})}} + \underbrace{V\left(E(Q|Y_{\text{obs}}, Y_{\text{mis}}) \mid Y_{\text{obs}}\right)}_{\substack{\text{Varianz d. Mittelwerte über alle MI-Sets} \\ \rightarrow \text{Between-Variance } (B)}}$$

Bestimmung der gepoolten Varianz von Parametern bei MI

Um die Varianz eines Parameters Q zu bestimmen (z.B. für Konfidenzintervalle), muss bei MI sowohl die (gemittelte) Varianz durch den Stichprobenfehler \bar{U} , als auch die Imputationsunsicherheit B mit berücksichtigt werden. Die Berechnung der gepoolten Varianz erfolgt durch die sog. **"Rubin-Regeln"** (s. n. Folie).

"Rubin's Rules" - Die Kombinationsregeln nach Rubin (Rubin, 1987)

Die Rubin-Regeln stellen eine allgemeine Formel dar, nach der die MI-Varianz eines Punktschätzers Q im konkreten Fall berechnet werden kann:

$$\begin{aligned}\hat{V} &= \overbrace{\left(\frac{1}{m} \sum_{\ell=1}^m \bar{U}_{\ell} \right)}^{\bar{U}} + \left(1 + \frac{1}{m} \right) \overbrace{\left(\frac{1}{m-1} \sum_{\ell=1}^m (\hat{Q}_{\ell} - \bar{Q})(\hat{Q}_{\ell} - \bar{Q})' \right)}^B \\ &= \bar{U} + \left(1 + \frac{1}{m} \right) B \\ &\Rightarrow \bar{U} + B \quad \text{as } m \rightarrow \infty\end{aligned}$$

...aber warum *drei* Variationsquellen?



Die Rubin-Formeln beziehen auch mit ein, dass immer nur eine finite Anzahl an Imputationssets generiert werden (\rightarrow Einbezug der **Simulationsvarianz**).

$$\begin{aligned}\hat{V} &= \overbrace{\left(\frac{1}{m} \sum_{\ell=1}^m \bar{U}_{\ell} \right)}^{\bar{U}} + \left(1 + \frac{1}{m} \right) \overbrace{\left(\frac{1}{m-1} \sum_{\ell=1}^m (\hat{Q}_{\ell} - \bar{Q})(\hat{Q}_{\ell} - \bar{Q})' \right)}^B \\ &= \bar{U} + \left(1 + \frac{1}{m} \right) B \\ &\Rightarrow \bar{U} + B \quad \text{as } m \rightarrow \infty\end{aligned}$$

→ Je größer m , desto **geringer fällt diese Komponente ins Gewicht**.
Insbesondere, wenn besonders genaue Varianzschätzungen notwendig sind,
empfiehlt sich daher eine **hohe Anzahl an Imputationssets**.

Ein niedriges m führt zu **Konfidenzintervallen**, die **etwas breiter** sind als
wenn $m \rightarrow \infty$ (niedrigere Effizienz). Dieser Unterschied ist in der Praxis
jedoch typischerweise **überschaubar**.

Metriken & Freiheitsgrade bei MI-Analysen

Metriken für Parameterschätzungen in MI (Van Buuren, 2018, Kapitel 2.3.5)

Relative Increase in Variance Due to Nonresponse (RIV): Relativer Anstieg der Varianz aufgrund der Imputationsunsicherheit (wenn $RIV > 1$: Imputationsvarianz größer als “echte” Varianz in Y):

$$r_Q = \frac{B_Q/m + B_Q}{\bar{U}_Q}$$

Fraction of Missing Information Due to Nonresponse (FMI): Anteil der Information über Q , die durch die Imputationsunsicherheit “verloren geht”:

$$\gamma_Q = \frac{(r_Q + 2)/(\nu_Q + 3)}{1 + r_Q}$$

Wobei ν (“nu”) für die Freiheitsgrade bei der Schätzung von Q steht.

Freiheitsgrade bei MI-Analysen (Van Buuren, 2018, Kapitel 2.3.6)

Die Freiheitsgrade eines Modells sind definiert als Anzahl der Beobachtungen nach Abzug der Modellparameter: $\nu = n - k$ (für Mittelwert z.B. $\nu = n - 1$).

Bei MI sind manche Elemente von n nur Schätzungen von Beobachtungen, daher muss ν dafür korrigiert werden (Rubin, 1987):

$$\nu_{(MI)} = (m - 1) \left(1 + \frac{1}{r^2} \right)$$
$$\lim_{r \rightarrow 0} \nu_{(MI)} = \infty \quad \text{sowie} \quad \lim_{r \rightarrow \infty} \nu_{(MI)} = m - 1$$

Diese Formel basiert auf der Annahme, dass die Freiheitsgrade des vollständigen Datensatzes (den MI zu schätzen versucht) unendlich groß sind! Diese Approximation ist aber erst sinnvoll, wenn ein relativ großes Sample vorliegt.

$$\nu_{\text{obs}} = \frac{\nu_{\text{com}} + 1}{\nu_{\text{com}} + 3} \nu_{\text{com}} \left(1 - \frac{r}{r + 1} \right)$$
$$\nu_{(\text{MI})}^* = \frac{\nu_{(\text{MI})} \nu_{\text{obs}}}{\nu_{(\text{MI})} + \nu_{\text{obs}}}$$

5.1.2 Researchers are often tempted to average the multiply imputed data, and analyze the averaged data as if it were complete. This method yields incorrect standard errors, confidence intervals and p -values, and thus should not be used if any form of statistical testing or uncertainty analysis is to be done on the imputed data. The reason is that the procedure ignores the between-imputation variability, and hence shares all the drawbacks of single imputation.

Dempster and Rubin (1983), this workflow is

... seductive because it can lull the user into the pleasurable state of believing that the data are complete after all.

The ensuing statistical analysis does not know which data are observed and which are missing, and treats all data values as real, which will underestimate the uncertainty of the parameters. The reported standard errors and

Referenzen

Rubin, D. B. (1987). *Multiple imputation for survey nonresponse*. New York: Wiley.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in *r*. *Journal of Statistical Software*, 45(1), 1–67.