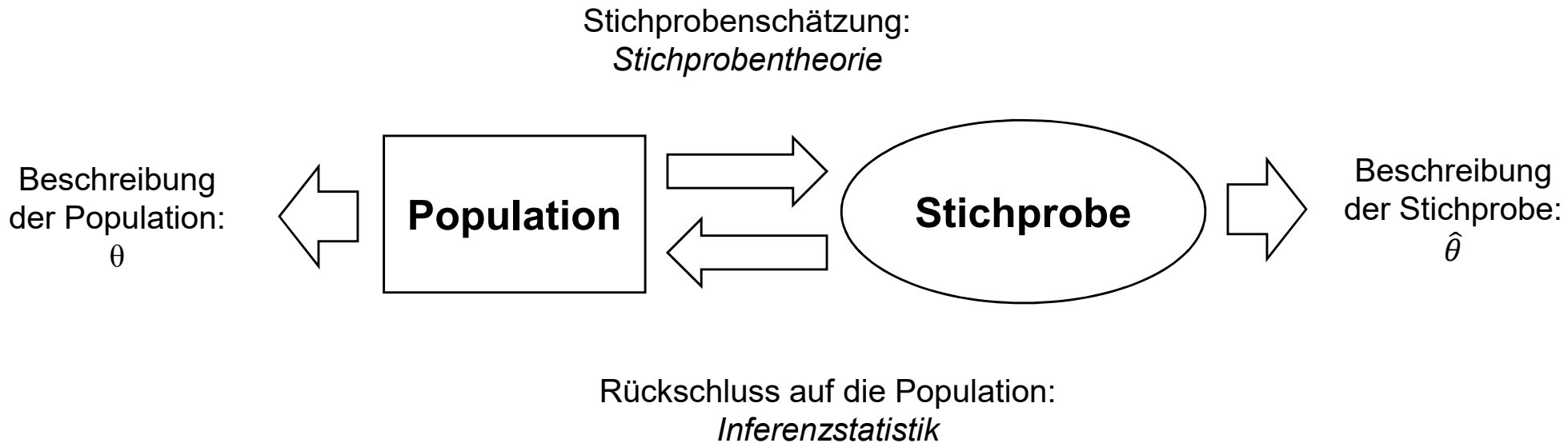
Ausblick

1. Arten von fehlenden Werten

2. Umgang mit fehlenden Werten

Populationsschätzung



- Das Ziel inferenzstatistischer Analysen ist die Schätzung unbekannter Populationskennwerte (-parameter) θ auf Basis von Stichprobendaten $\hat{\theta}$:
  - univariat (z.B.: Mittelwert, Median)
  - bivariat (z.B.: Regressionsgewicht)
  - multivariat (z.B.: Partialkorrelation)

Konsequenzen fehlender Werte

- Welche Auswirkung haben fehlende Werte in Stichprobendaten auf die Schätzung unbekannter Populationsparameter?
- *Beispiel*: Wie zufrieden sind die Österreicher mit ihrem Leben? Gesucht ist eine möglichst unverzerrte Schätzung von Mittelwert und Varianz der Lebenszufriedenheit der österreichischen Bevölkerung.
  - Stichprobe: repräsentative Erhebung von Bürgerinnen und Bürgern Österreichs (ab 16 Jahren)
  - Item: *Wie zufrieden sind Sie im Allgemeine mit Ihrem Leben?*
    (Antwortskala: 0 = *sehr unzufrieden* bis 10 = *sehr zufrieden*)

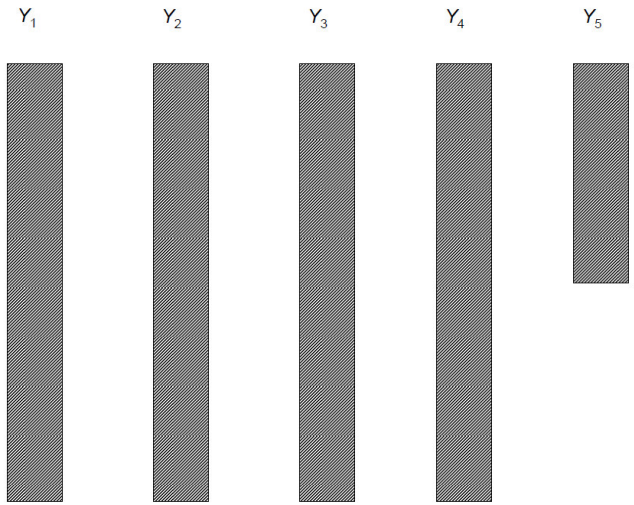
	Vollständig	Zufälliger Ausfall	Fehlende Werte bei Arbeitslosen	Fehlende Werte bei niedriger Lebenszufriedenheit
N	1,000	900	900	900
Fehlend	0%	10%	10%	10%
M	7.34	7.37	7.52	7.74
Md	8.00	8.00	8.00	8.00
Var	4.15	4.10	3.80	3.00

Arten von Non-Response

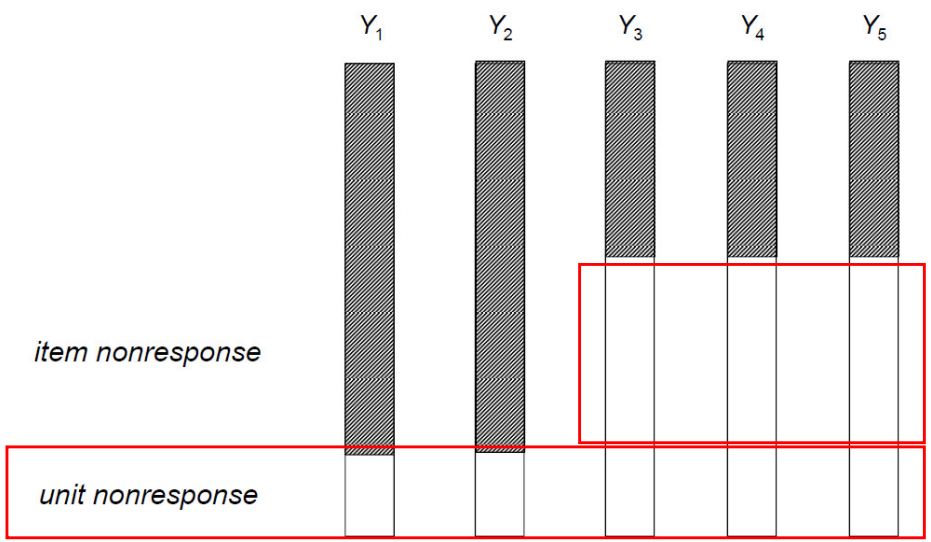
Non-response (=Ausfälle):

- Beschreibt den Umstand, dass bei einem Teil der Personen, die zur Stichprobe gehören, aus verschiedenen Gründen keine Befragung stattfindet. Die Ausfälle können sich dabei auf die gesamte Befragung (=unit non-response) oder nur einzelne Fragen (=item non-response) erstrecken.

Univariate non-response



Multivariate non-response



Response continuum model:

0% Wahrscheinlichkeit Fragen zu beantworten	Geringe Wahrscheinlichkeit Fragen zu beantworten	Hohe Wahrscheinlichkeit Fragen zu beantworten	100% Wahrscheinlichkeit Fragen zu beantworten

Gründe für Non-response

Unit non-response	Item non-response
• Mangelnde Erreichbarkeit des Befragten im Erhebungszeitraum (es kommt kein Kontakt zustande) • Befragte möchte nicht an der Befragung teilnehmen (Teilnahme-verweigerung) • (temporäre) Unfähigkeit des Befragten an der Befragung teilzunehmen (z.B. aufgrund von Sprachproblemen oder Krankheit)	• unzureichendes Verständnis einzelner Fragen • Befragte/r möchte keine Antwort geben (z.B. Antwortverweigerung bei sensitiven Themen) • Unfähigkeit des Befragten eine Antwort zu geben (z.B. weil er/sie meint nicht über die nötigen Informationen zu verfügen) • Befragte nimmt Item nicht wahr oder hat keine Lust sich damit zu beschäftigen (Unachtsamkeit, Motivationsmangel)

Gründe für Unit Non-response

- **Mangelnde Erreichbarkeit**:
  - fehlende bzw. eingeschränkte Zugangsmöglichkeiten (z.B. bei Nicht-Annahme von unbekannten Telefonanrufen, Wohnung in unzugänglichem Gebiet)
  - fehlende Anwesenheit (abhängig von soziodemographischen Merkmalen)
  - fehlende Häufigkeit oder falsche Art der Kontaktversuche
- **Teilnahmeverweigerung**:
  - Merkmalen der Befragten (z.B. bei Männern höher)
  - Merkmalen und Vorgehen der Interviewer
  - Interaktion zwischen Interviewer und Befragten
  - Design der Umfrage (z.B. Incentivierung, Mode)
- Interviews können z.B. **nicht durchgeführt** werden bei:
  - Sprachproblemen
  - Krankheit/Behinderung
  - Schreib- und Leseschwächen (bei schriftl. Befragung)
  - Fehlen von Informationen für die Beantwortung der Fragen

PRÜFUNGSRELEVANT : BEGRIFFLICHKEITEN KENNEN

Mechanismen für das Auftreten fehlender Werte

- Nach Rubin (1976) können drei Arten von fehlenden Werten (**missing values**) unterschieden werden:

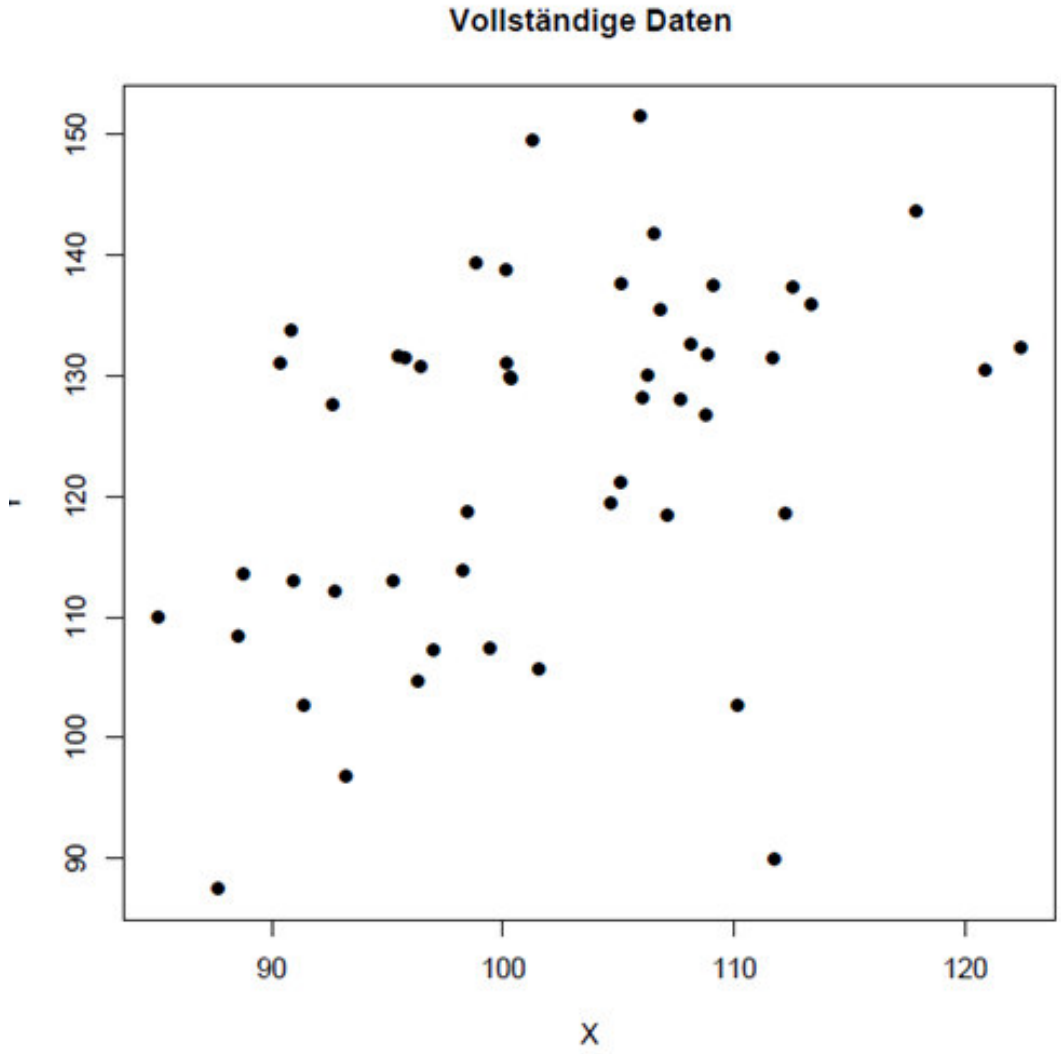
Missingverhalten komplett unabhängig

Missingverhalten von 3. Variable abhängig (Bsp. ob ich Einkommen angeben ist vom Alter abhängig

Wahrscheinlichkeit eine Antwort zu geben ist von der Variable abhängig

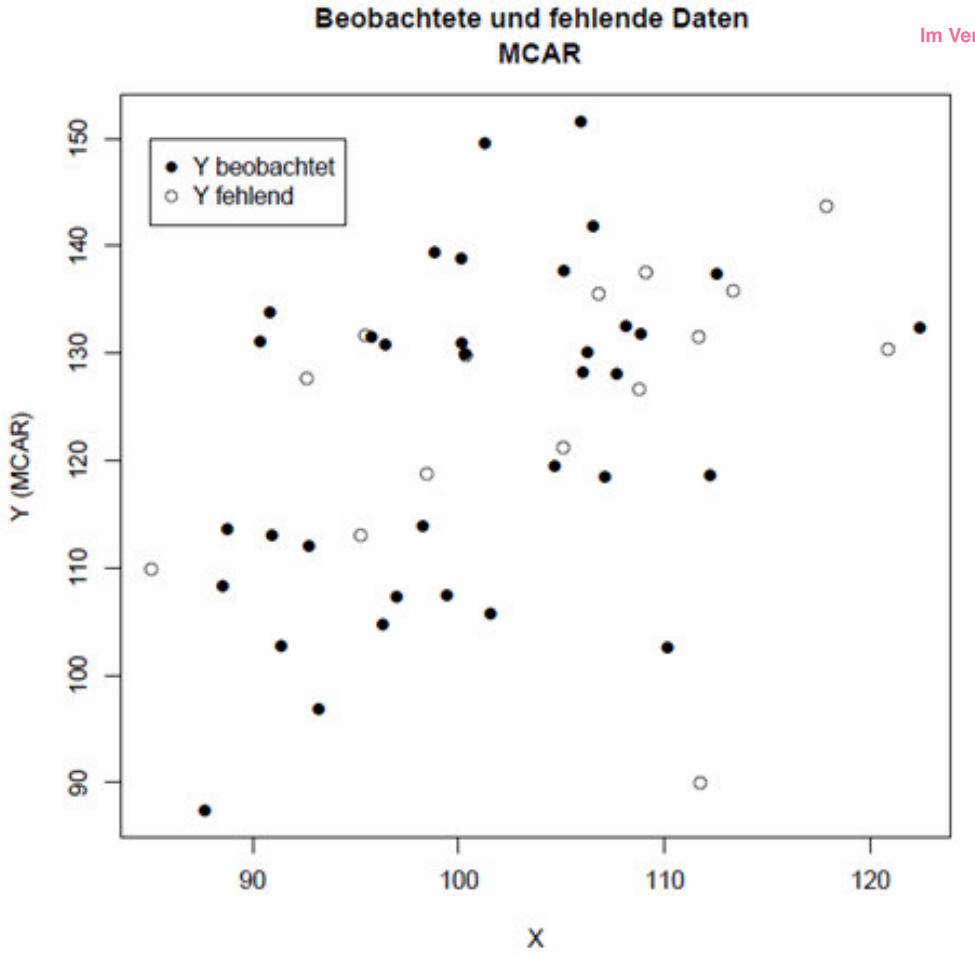
Missing Completely at Random (MCAR)	Missing at Random (MAR)	Missing Not at Random (MNAR)
• Die fehlenden Werte stellen eine Zufallsstichprobe aus der Gesamtstichprobe dar. • Die Wahrscheinlichkeit eines non-response korreliert nicht mit Drittvariablen. • Die Wahrscheinlichkeit eines non-response korreliert nicht mit der eigentlich interessierenden (kritischen) Variable.	• Die Wahrscheinlichkeit eines non-response korreliert mit einer oder mehreren Drittvariablen. • Die Wahrscheinlichkeit eines non-response korreliert nicht mit der eigentlich interessierenden (kritischen) Variable. • Nach Kontrolle der Drittvariable ist die Wahrscheinlichkeit eines non-response zufällig.	• Die Wahrscheinlichkeit eines non-response korreliert mit der eigentlich interessierenden (kritischen) Variablen. • Die Korrelation ist selbst nach Kontrolle von Drittvariablen zu beobachten.
• *Beispiel*: Fehlende Angabe des Einkommens hängt weder vom Einkommen des Befragten noch von einer anderen Variable ab.	• *Beispiel*: Jüngere Befragte geben ihr Einkommen seltener an, unabhängig von der Höhe des Einkommen.	• *Beispiel*: Insbesondere wohlhabende Befragte verweigern eine Auskunft über ihr Einkommen.

Beispiel: Vollständige Daten



	Vollständig
N	50
$M(X)$	101.9
$M(Y)$	123.9
$SD(X)$	8.9
$SD(Y)$	14.8
$Cov(X,Y)$	54.4
$r(X,Y)$	.41

Beispiel: Missing Completely at Random (MCAR)



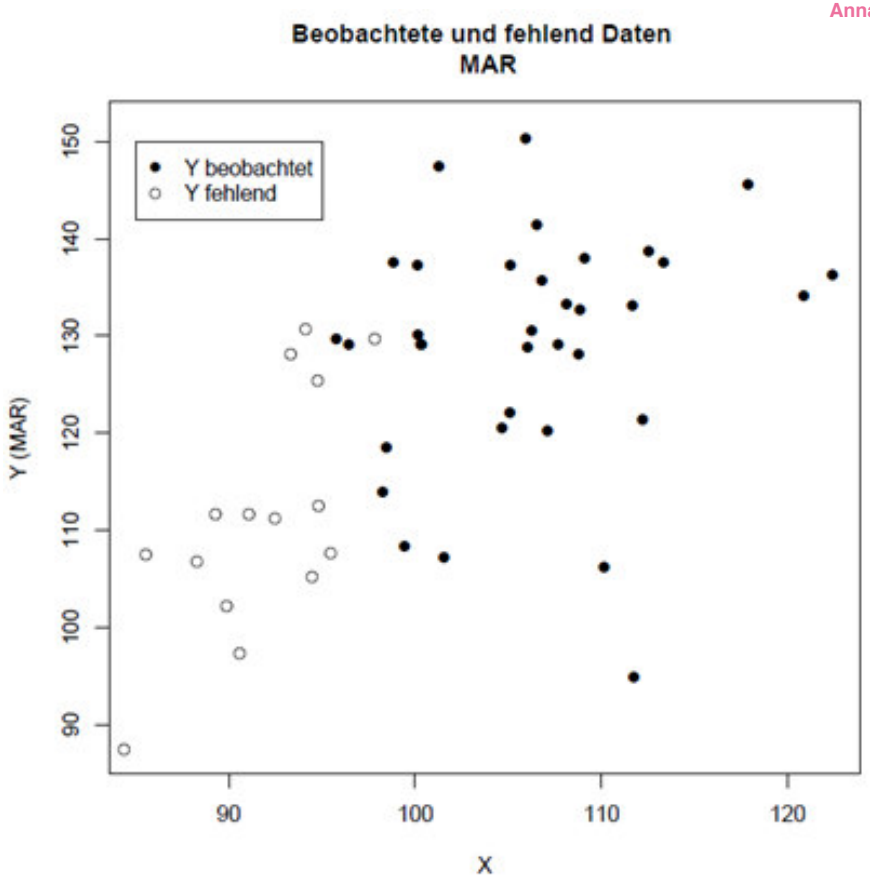
Im Vergleich zu allen Daten kein großer Unterschied. Y Werte ähnlihe

	Vollständig	MCAR*
N	50	35
$M(X)$	101.9	100.7
$M(Y)$	123.9	122.8
$SD(X)$	8.9	8.1
$SD(Y)$	14.8	15.5
$Cov(X,Y)$	54.4	54.4
$r(X,Y)$	.41	.43

* fallweiser Ausschluss

Der Ausfall von Y ist komplett zufällig und hängt weder mit Drittvariablen noch mit der Höhe von Y ab (hier: 30% fehlende Werte in Y).

Beispiel: Missing at Random (MAR)



Annahme: Ausprägung der Variable Y ist von zusätzlicher Variable x abhängig

	Vollständig	MCAR*	MAR*
N	50	35	35
$M(X)$	101.9	100.7	106.3
$M(Y)$	123.9	122.8	127.6
$SD(X)$	8.9	8.1	6.6
$SD(Y)$	14.8	15.5	13.5
$Cov(X,Y)$	54.4	54.4	11.4
$r(X,Y)$	.41	.43	.13

* fallweiser Ausschluss

Jetzt verschwindet die Korrelation. Die Punktschätzung wurde deutlich verzerrt

Der Ausfall von Y hängt von X ab, aber nicht mit der Höhe von Y ab (hier: 30% fehlende Werte in Y).

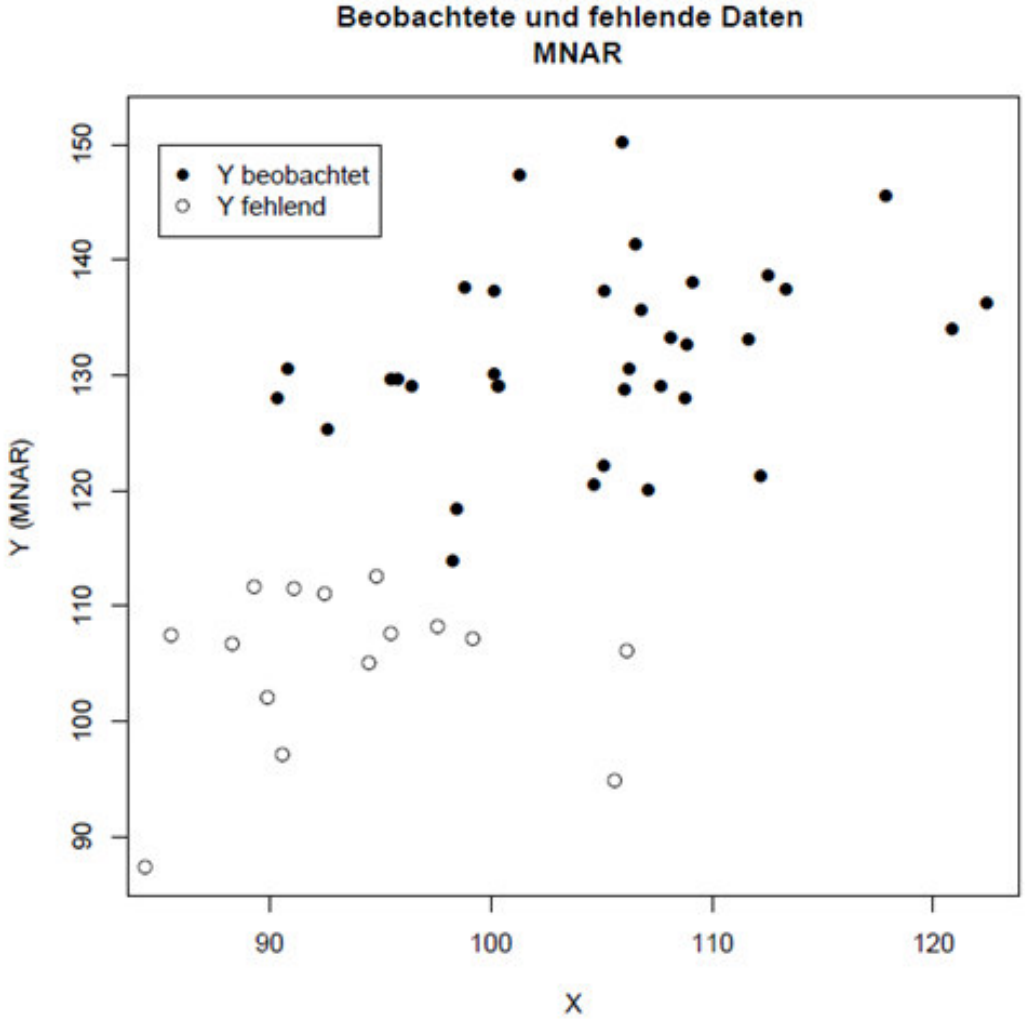
Identifizierung von MCAR versus MAR

- **Zielsetzung**: Ist die Annahme von MCAR gerechtfertigt?
- **Vorgehen**: Prüfung, ob sich verschiedene Drittvariable zwischen Pbn mit bzw. ohne fehlende Werte signifikant von einander unterscheiden.
- **Univariate *t*-Tests** für unabhängige Stichproben für jede Variable:
  - Vor- und Nachteile: + einfache Berechnung, - ignoriert Abhängigkeiten zwischen Drittvariablen, - Inflation des Fehlers 1. Art aufgrund vieler Tests, - häufig kleine Gruppengröße
- **Little‘s (1988) MCAR Test** (= multivariater Tests):
  - Globaler Test für all Variablen eines Datensatzes
  - Vergleicht Mittelwerte zwischen Gruppen von Pbn, die unterschiedliche Muster von fehlenden Werten aufweisen
  - Testet die H_0, dass fehlende Werte zufällig verteilt sind
  - Nachteile: - zeigt nicht, welche Variable für die MCAR-Verletzung verantwortlich ist (= globaler Test), - geringe Teststärke (insb. bei wenigen Variablen mit MCAR, geringer Korrelation zwischen Daten und fehlenden Werten, oder Vorliegen von NMAR)

Beispiel: Missing at Not at Random (MNAR)

Wahrscheinlichkeit das ein Wert fehlt hängt von Variable selbst ab

Mittelwert Y ist überschätzt, weil die niedrigen Werte fehlen. Varianz (SD) von Y ist kleiner, Korrelation ist deutlich zu klein (.11)



	Vollständig	MCAR*	MAR*	MNAR*
N	50	35	35	35
$M(X)$	101.9	100.7	106.3	104.7
$M(Y)$	123.9	122.8	127.6	131.6
$SD(X)$	8.9	8.1	6.6	7.8
$SD(Y)$	14.8	15.5	13.5	8.4
$Cov(X,Y)$	54.4	54.4	11.4	7.5
$r(X,Y)$	.41	.43	.13	.11

* fallweiser Ausschluss

Der Ausfall von Y hängt von der Höhe von Y selbst ab (hier: 30% fehlende Werte in Y).

Identifizierung von MAR versus MNAR

Es gibt keinen Test um MNAR von anderen Gründen warum ein Wert fehlt, abzugrenzen

Wichtig

Es gibt **keine Tests**, um MNAR zu prüfen.

- Daher wird empfohlen, anhand möglichst vieler Drittvariable MAR zu belegen und im Rahmen der Datenanalyse zu berücksichtigen, um dadurch einen möglichen MNAR-Mechanismus möglichst gering zu halten.

Welche untersch. Missingmechanismen gibt es?
Was für Auswirkungen haben diese auf die Anwenung statistischer Verfahren

Ausblick

1. Arten von fehlenden Werten
2. **Umgang mit fehlenden Werten**

Ausschluss von Fällen

Problem: Schliesse ich aus, wird meine Stichprobe kleiner und meine Varianz ist geringer
. Varianzeischränkung

Listenweiser Fallausschluss (complete-case analysis)	Paarweiser Fallausschluss (available-case analysis)
• Nur Personen mit vollständigen Datenmatrizen auf allen interessierenden Variablen werden einbezogen • Voreinstellung in den meisten Softwarepaketen • Geringe Power durch Reduktion der Stichprobengröße	• Es werden alle für die spezifische Analyse verfügbaren Personen berücksichtigt • Höhere Power als bei listenweisem Fallausschluss • Kann zu statistischen Problemen führen (z.B. nicht-positiv definite Korrelationsmatrizen) • Es ist unklar wie Standardfehler (z.B. in Regressionsanalysen) berechnet werden sollen, da unterschiedliche N für die Variablen vorliegen.

Führt zu verzerrten Schätzungen, wenn MCAR nicht gegeben ist!

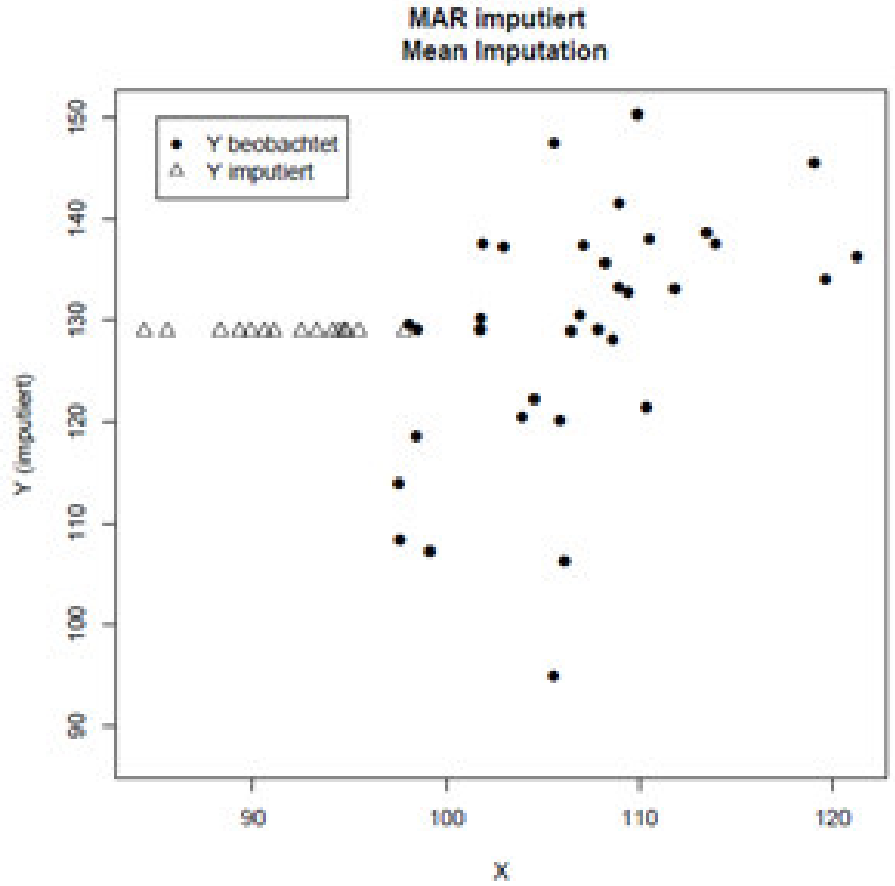
Ersetzen fehlender Werte

- **Prinzip**: Jeder fehlende Wert wird durch einen oder mehrere möglichst „plausible“ Werte ersetzt
- **Konsequenz**:
  - Es wird ein vollständiger Datensatz generiert.
  - Dadurch ergeben sich effizientere Schätzung (= höhere Teststärke).
  - Dadurch können alle beobachteten Informationen genutzt werden (= präzisere Schätzung).
  - ABER: ggf. Varianzeinschränkung oder verzerrte Parameterschätzungen

Single Imputation (Ersetzen durch einen Wert)	Multiple Imputation (Ersetzen durch eine Verteilung von Werten)
• Mittelwertsersetzung • Deterministische Regressionsimputation • Stochastische Regressionsimputation	• Joint Modeling (gemeinsame Verteilung) • Fully conditional modeling (Folge von bedingten Verteilungen)

Konsequenz für die Statistik: Wir haben jetzt wieder n=50.
Varianz ist kleiner gewerden. Wenn wir die Varianz unterschätzen verschätzen wir uns in der Korrelation noch stärker.
Mittelwerte Ersetzen ist nicht ganz fir Clevere Idee wegen den Verzerrungen.

Beispiel: Mittelwertsersetzung



Jeder fehlende Wert wird durch den Mittelwert der Variable ersetzt

	Voll-ständig	MAR*	Imput.: M
N	50	35	50
$M(X)$	101.9	106.3	101.9
$M(Y)$	123.9	127.6	127.9
$SD(X)$	8.9	6.6	8.9
$SD(Y)$	14.8	13.5	11.3
$Cov(X,Y)$	54.4	11.4	7.9
$r(X,Y)$	.41	.13	.08

* fallweiser Ausschluss

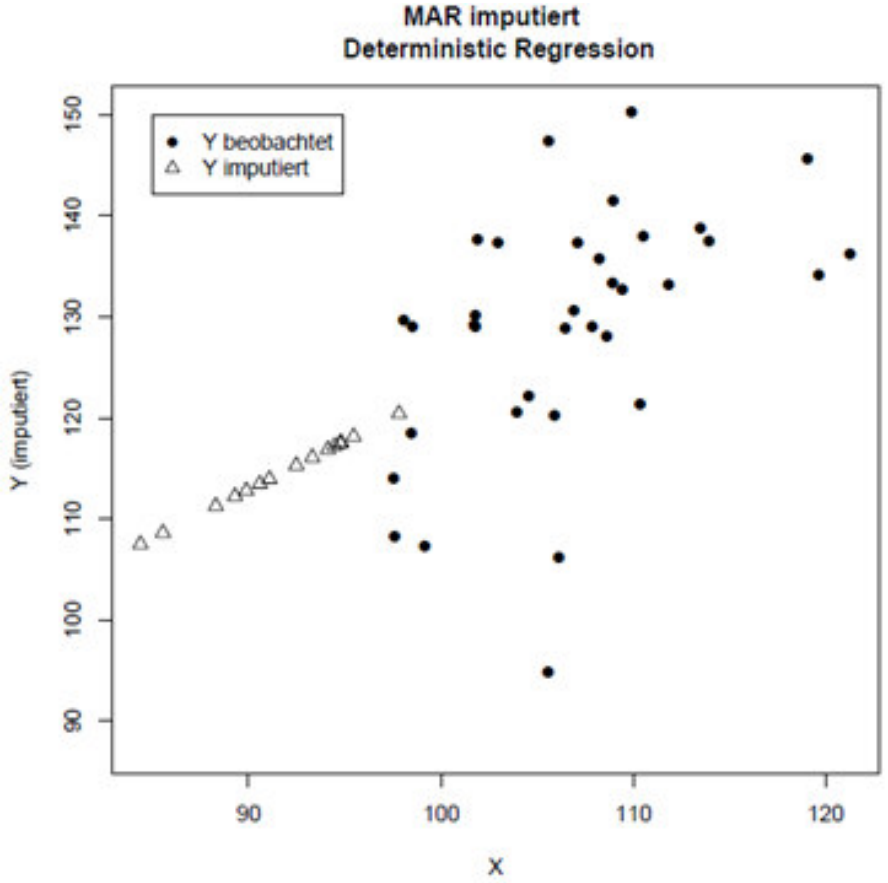
- Fehlende Werte werden durch den Mittelwert von Y ersetzt.
- Führt zu verzerrten Schätzungen des Mittelwerts und zur Unterschätzung der Varianz, selbst bei MCAR.

Deterministische Regressionsimputation

- Fehlende Werte auf der kritischen Variable (Y) werden auf Basis eines multiplen Regressionsmodells ersetzt.
- Die Prädiktoren sollten hierbei theoretisch fundierte Variablen sein, die den Wert der kritischen Variable Y vorhersagen.
- **Ablauf**:
  1. Regression von Y auf $X_1, ..., X_m$ auf Basis der vollständigen Daten
  2. Einsetzen in die Regressionsgleichung, um Vorhersagen der fehlenden Werte in Y zu treffen.
  3. Imputieren der fehlenden Werte auf Y durch das geschätzte $\hat{Y}$
- **Nachteile**:
  - Führt zu verzerrten Schätzungen von Varianzen und Kovarianzen, die unter MCAR korrigert werden können (vgl. Enders, 2010).
  - Überschätzt den Zusammenhang zwischen Y und $X_1, ..., X_m$.

Hier ist der Mittelwert etwa gleich geblieben, Varianz ist immer noch eingeschränkt und unterschätzt. Folglich ist die Korrelation etwas besser aber immer noch unterschätzt

Beispiel: Deterministische Regressionsimputation



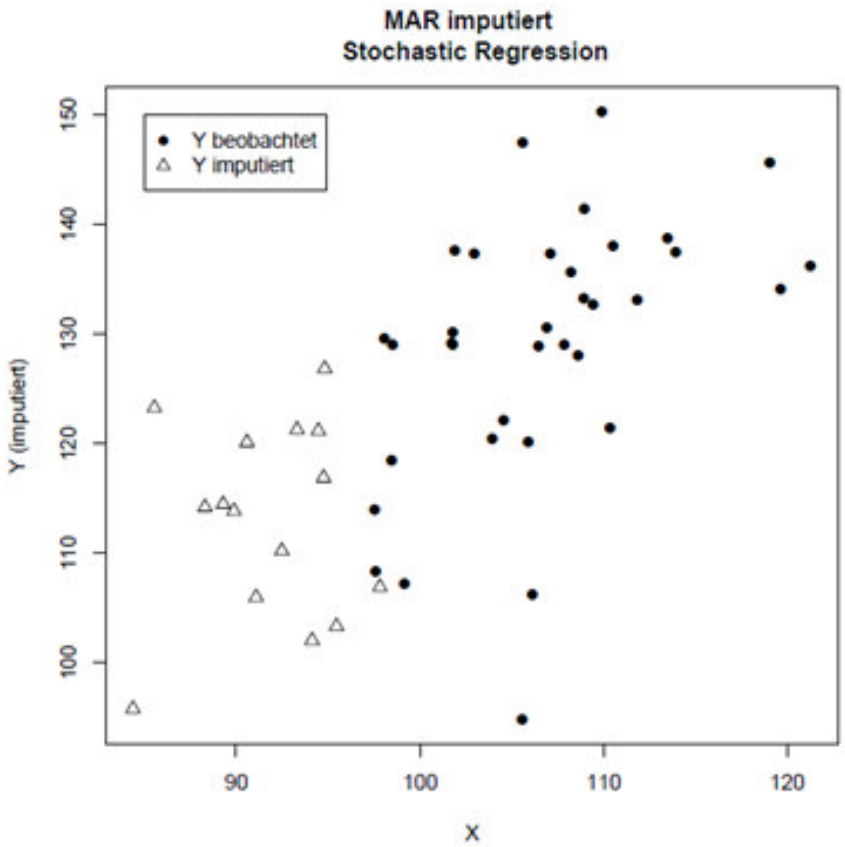
	Voll-ständig	MAR*	Imput.: M	Imput.: Det. Regr.
N	50	35	50	50
$M(X)$	101.9	106.3	101.9	101.9
$M(Y)$	123.9	127.6	127.9	126.4
$SD(X)$	8.9	6.6	8.9	8.9
$SD(Y)$	14.8	13.5	11.3	11.4
$Cov(X,Y)$	54.4	11.4	7.9	20.9
$r(X,Y)$	.41	.13	.08	.21

* fallweiser Ausschluss

- Fehlende Werte auf Y werden durch Regression von Y auf X ersetzt.
- Für jeden einzelnen Fall beste Vorhersage als imputierter Wert liefert keine optimalen Parameterschätzungen.

Beispiel: Stochastische Regressionsimputation

Mittelwert ist besser geworden. Liegt wahr am vollständigen Datensatz. SD ist bei 13.4, Korrelation zwischen X und Y die uns eigentlich interessiert nähert sich stark dem wahren Datensatz an



	Voll-ständig	MAR*	Imput.: M	Imput.: Det. Regr.	Imput.: Sto.Regr.
N	50	35	50	50	50
$M(X)$	101.9	106.3	101.9	101.9	101.9
$M(Y)$	123.9	127.6	127.9	126.4	124.9
$SD(X)$	8.9	6.6	8.9	8.9	8.9
$SD(Y)$	14.8	13.5	11.3	11.4	13.4
$Cov(X,Y)$	54.4	11.4	7.9	20.9	37.9
$r(X,Y)$	.41	.13	.08	.21	.32

* fallweiser Ausschluss

- Fehlende Werte auf Y werden durch Regression von Y auf X sowie einem Fehler (= zufällige Ziehung aus Normalverteilung oder empirischen Residuen) ersetzt.
- Für jeden einzelnen Fall bedingt zufällige Vorhersage als imputierter Wert liefert „bessere“ Parameterschätzungen, die unter MAR unverzerrt sind.

Zusammenfassung Single Imputation

- **Single Imputation** führt zu verzerrten Parameterschätzung. Diese Verzerrungen fallen für die verschiedenen Ansätze jedoch unterschiedlich hoch aus:

Stochastische Regressionsimputation

<

Deterministische Regressionsimputation

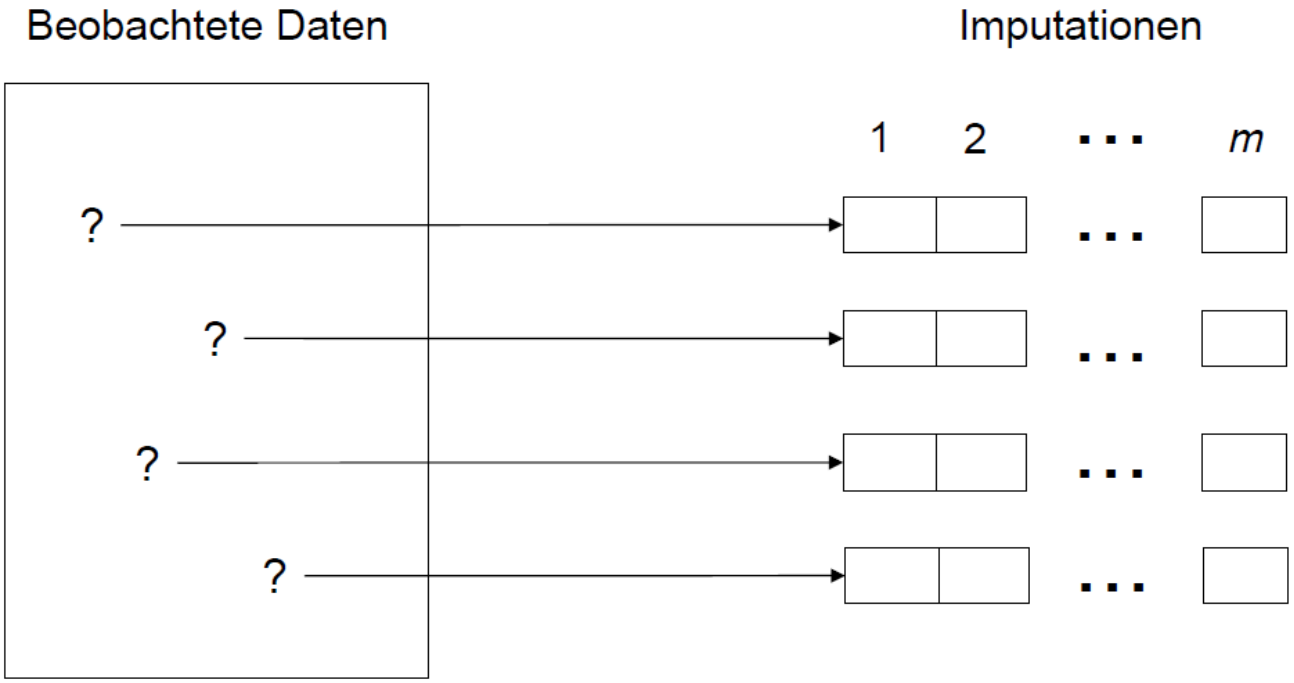
<

Mittelwertsimputation

- **Nachteile**:
  - Berücksichtigt nur ungenügend die Unsicherheit mit der fehlende Werte ersetzt werden
  - Unterschätzung des Standardfehlers (erhöhtes Risiko Alphafehler zu begehen)

Multiple Imputation

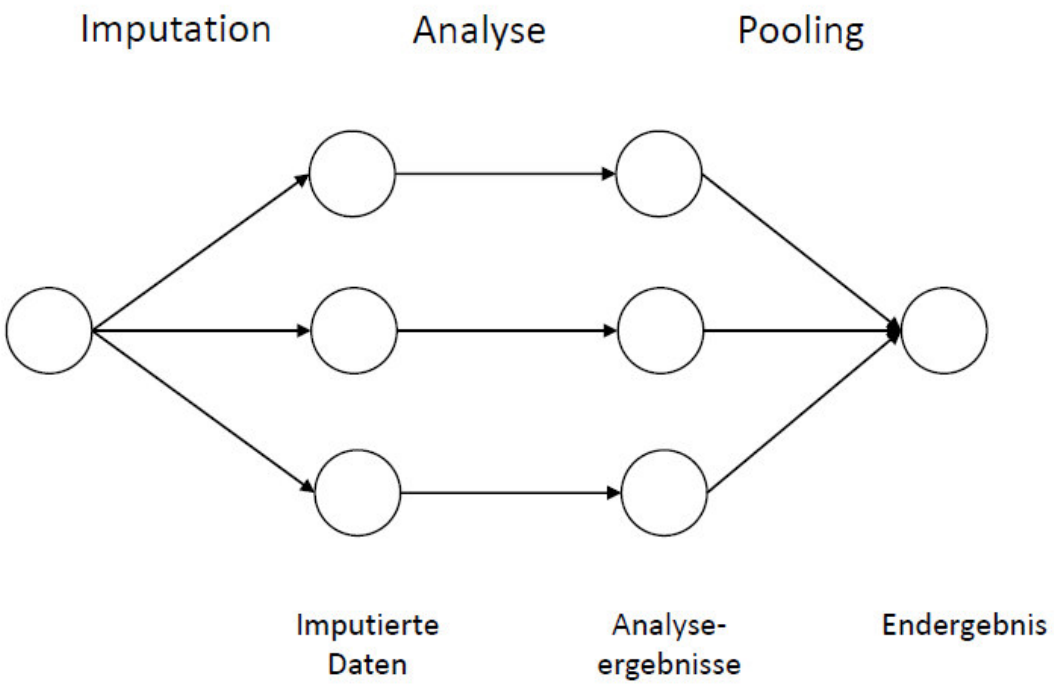
- Bei der **multiplen Imputation** werden fehlende Werte durch mehrere unterschiedliche Werte ersetzt, um die Unsicherheit bei der Ersetzung der fehlenden Werte zu berücksichtigen:
  - Varianz innerhalb der Imputationen
  - Varianz zwischen den Imputationen
- Es werden mehrere vollständige Datensätze erzeugt.



Schritte bei einer multiplen Imputation

- **Ablauf**:
  1. Für jeden fehlenden Wert werden unter Berücksichtigung von kausalen Annahmen mehrere Ersetzungen (z.B. über eine stochastische Regressionsimputation) vorgenommen und dadurch insgesamt m (z.B. 50) vollständige Datensätze erzeugt.

     Dabei sind nicht nur Variablen des Analysemodells zu berücksichtigen; es sollen alle Variablen, die zur Vorhersage beitragen können, herangezogen werden.
  2. Jeder Datensatz wird mit Standardverfahren analysiert. Die geschätzten Parameter auf Basis eines Datensatzes werden als fehlerbehaftet angenommen, die einer Verteilung unterliegen.
  3. Die Ergebnisse aus den m Analysen werden zusammengefasst.



Zusammenfassen der Ergebnisse

Die Ergebnisse aus m Analysen multiple imputierter Daten lassen sich nach Rubin (1987) zusammenfassen:

- Die **Punktschätzung** eines Parameters θ ergibt sich als Mittelwert der m Parameter:

$$\theta = \frac{1}{m}\sum_{t=1}^{m}\hat{\theta}_t$$

- Der **Standardfehler** SE für einen Parameter ergibt sich additiv aus …
  - der Varianz innerhalb der Imputationen und
  - der Varianz zwischen den Imputationen.

Sensibilisiert sein, dass Werte Probleme darstellen können, ganz kurz grobe Idee der Single Imputation Verfahren sagen können und die Logik einer multiplen Imputation widergeben können

Wichtig

Standardfehler aus multipler Imputation sind **größer** als bei nicht-imputierten Daten, weil sie auch die Unsicherheit aufgrund fehlender Werte berücksichtigen.

Zitierte Quellen

- 🕮 Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.
- 🕮 Rubin, D. B. (1987). *Multiple imputation for non response in surveys*. Hoboken, NJ: Wiley.