

Wissenschaftliches Arbeiten und Forschungsmethoden

Einheit 6: Samplingplan

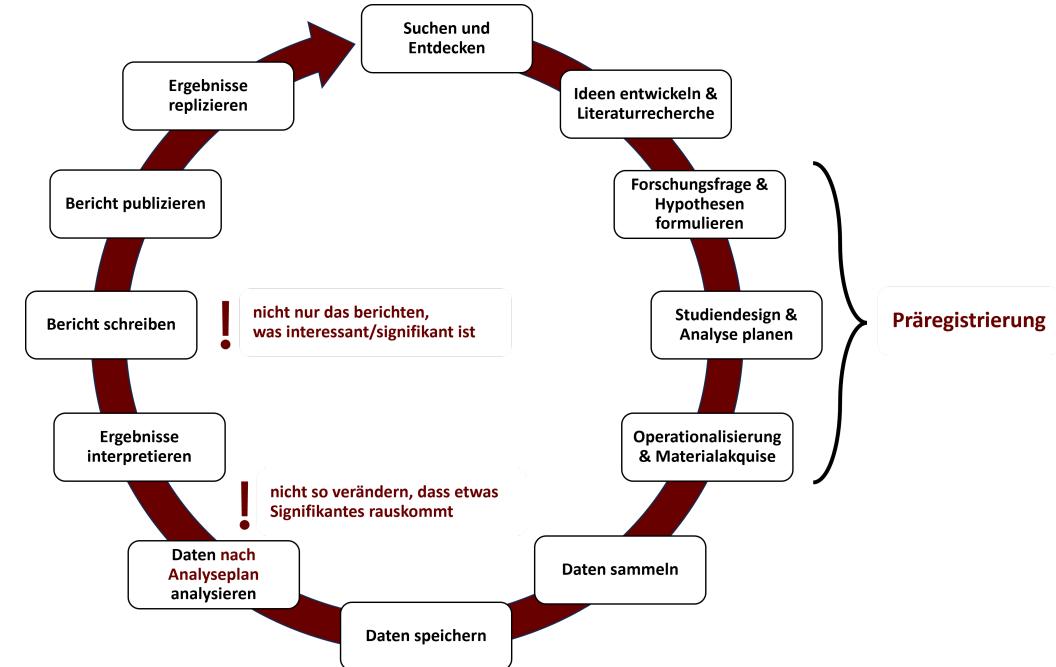
22.05.2024 | Prof. Dr. Stephan Goerigk

Sampling Plan (Rekrutierungsplanung)

Sampling Plan im Forschungsprozess

Was gehört zur Studienplanung?

1. Theoriearbeit und Literaturrecherche → Einheit 2 und 3
 2. Hypothesenformulierung → Einheit 3
 3. Design Plan → Einheit 3
 4. Variablen (Auswahl Messinstrumente und Variablenrollen)
→ Einheit 4
 5. Statistischer Analyseplan → letzte Sitzung
 6. Sampling plan (Rekrutierungsplan) → heutige Sitzung
- Studienplanung in **Präregistrierung** festhalten



Sampling Plan (Rekrutierungsplanung)

Elemente

- Falls zutreffend: Beschreibung der Regeln zur Vorauswahl von Teilnehmer:innen
- Beschreibung der Datenerhebung (Wo? Von wem? Wie? Wann?)
- Rechtfertigung der geplanten Stichprobengröße (→ Poweranalyse)
- Beschreibung der Regel zur Beendigung der Datenerhebung
- ggf. Umgang mit Personen, die die Studie abbrechen ("drop out")
- ggf. Wenn-dann-Regeln zum Umgang mit zu kleiner Stichprobe nach Ausschluss

Auswahl der Untersuchungsobjekte

- Frage: an welchen bzw. an wie vielen Untersuchungsobjekten werden die Variablen erhoben? → Untersuchungsteilnehmer, deren Beobachtung oder Beschreibung interessante Hypothesen versprechen

Wichtigstes Kriterium: Repräsentativität der Stichprobe

- Repräsentativität = Stichprobe muss in ihrer Zusammensetzung der Population möglichst stark ähneln
- Stichprobe ist (merkmals)spezifisch repräsentativ, wenn ihre Zusammensetzung hinsichtlich relevanter Merkmale der Populationszusammensetzung entspricht
- Stichprobe ist global repräsentativ, wenn ihre Zusammensetzung in nahezu allen Merkmalen der Populationszusammensetzung entspricht
- große Stichprobe garantiert nicht Repräsentativität (gilt nur bei unverzerrter Auswahl)
- beste Gewähr für größtmögliche globale Repräsentativität bietet die Zufallsstichprobe

Sampling Plan (Rekrutierungsplanung)

Auswahl der Untersuchungsobjekte

Einschlusskriterien:

- globale Population aller Menschen ist nicht für jede Fragestellung geeignet
- Selektion zur Erhöhung der Auftretenswahrscheinlichkeit relevanter Merkmale (z.B. bestimmte Diagnose, bestimmte Altersklasse...)

Ausschlusskriterien:

- Manche Charakteristika machen Untersuchung einzelner Personen inhaltlich/ethisch unmöglich
- Ausschlusskriterien müssen gut überlegt und genau dokumentiert werden

→ Für Gruppen mit eigenem Thema gibt es ein Einschlusskriterium, welches nicht inhaltlich begründet ist: Studierender dieser Vorlesung sein. Fehlende Repräsentativität ist dann Teil der Diskussion.

→ Für Gruppen mit Interventionsthema gibt es als Einschlusskriterium, dass die Versuchspersonen in einer festen, monogamen Partnerschaft sein sollen.

Sampling Plan (Rekrutierungsplanung)

Anwerbung von Untersuchungsteilnehmern:

- Wo soll rekrutiert werden?
- Mit welchen Mitteln?
- Wird es eine Vergütung geben?
 - Bei uns: Rückmeldung über eigene Antworten
- Ist die Teilnahme Teil eines erweiterten (Forschungs-/Behandlungs-)Programms?

Sampling Plan (Rekrutierungsplanung)

Determinanten der freiwilligen Untersuchungsteilnahme (Bortz und Döring, 2012)

Freiwillige Untersuchungsteilnehmer ...

- ...verfügen über eine bessere schulische Ausbildung als Verweigerer (bessere Notendurchschnitte).
- ...schätzen ihren eigenen sozialen Status höher ein als Verweigerer.
- ...benötigen mehr soziale Anerkennung als Verweigerer.
- ...sind geselliger als Verweigerer.
- ...sind weniger autoritär als Verweigerer.
- ...haben eine geringere Tendenz zu konformem Verhalten als Verweigerer.
- ...geben sich in Untersuchungen über geschlechtsspezifisches Verhalten unkonventioneller.
- ...verfügen auf Basis der meisten Untersuchungsergebnisse über eine höhere Intelligenz.

Im Allgemeinen sind weibliche Personen eher zur freiwilligen Untersuchungsteilnahme bereit als männliche Personen.

Siehe auch Rezeption einer aktuelleren Studie: <https://twitter.com/rubenarslan/status/1700049223889101230>: Freiwillige gehen eher wählen und antworten eher "Ja" auf Fragen

Sampling Plan (Rekrutierungsplanung)

Zeitpunkt der Registrierung

In einer Präregistrierung wird der Zeitpunkt der Registrierung konkret dokumentiert:

- **Registrierung vor der Sammlung der Daten:** Daten noch nicht erhoben, erstellt oder realisiert worden
- **Registrierung vor jeglicher menschlicher Beobachtung der Daten:** Daten vorhanden, wurden aber noch von niemandem quantifiziert, erstellt, beobachtet oder gemeldet
- **Registrierung vor dem Zugriff auf die Daten:** Daten vorhanden, aber Sie oder Ihre Mitarbeiter haben noch keinen Zugriff darauf
- **Registrierung vor der Analyse der Daten:** Daten vorhanden und Sie haben darauf zugegriffen, aber es wurde noch keine Analyse im Zusammenhang mit dem Forschungsplan durchgeführt
- **Registrierung nach der Analyse der Daten:** Sie haben auf einen Teil der für den Forschungsplan relevanten Daten zugegriffen und diese analysiert

Wenn für eine Studie auf einem bereits bestehenden Datensatz Berechnungen durchgeführt werden (Analysen auf vorab-existierenden Daten), muss dies auch durchdacht und dokumentiert werden → siehe "Preregistration of secondary data analysis: A template and tutorial":

<https://research.tilburguniversity.edu/en/publications/preregistration-of-secondary-data-analysis-a-template-and-tutorial>

Sampling Plan (Rekrutierungsplanung)

Größe der Stichprobe

- Die Festlegung des Stichprobenumfanges sollte ebenfalls in der Planungsphase erfolgen
- Verbindliche Angaben zum nötigen Stichprobenumfang für eine bestimmte Analyse nur möglich, wenn eine hypothesenprüfende Untersuchung mit vorgegebener Effektgröße geplant wird → **a priori Poweranalyse**
- Durchführung z.B. in R oder freies Programm GPower ([Link zu Erklärvideo für G*Power](#))
- Für die Größe von Stichproben, mit denen keine spezifischen Hypothesen geprüft werden oder keine Effektgröße schätzbar ist, gibt es keine genauen Richtlinien ("more is better, but be mindful of resources")
- Falls die Stichprobengröße aus anderen Gründen festgelegt ist (wie bei den Gruppen mit eigenem Thema), kann man eine "**(post hoc) Sensitivitäts-Poweranalyse**" machen, um abzuschätzen wie hoch die statistische Power für die Aufdeckung eines Effekts mit dieser Stichprobengröße ist (wichtig: eine post hoc Analyse auf dem *gefundenen* Effekt aus der eigenen Studie macht keinen Sinn, nur eine auf dem erwarteten Effekt)

Sampling Plan (Rekrutierungsplanung)

Größe der Stichprobe

- Ziel der a priori Poweranalyse: Bestimmung einer Stichprobengröße, die bei angenommender Effektgröße eine statistisch-signifikantes Ergebnis ermöglicht
- Zusammenspiel aus Signifikanzniveau (Wahrscheinlichkeit für Fehler 1. Art) und Power (1 - Wahrscheinlichkeit für Fehler 2. Art)
- Poweranalyse richtet sich nach der Hauptfragestellung (primärer zu überprüfender Hypothese); bei mehreren Primärvorschlägen nach dem kleinsten erwarteten Effekt

Relevante Parameter:

- **Effektstärke** (aus Vorstudien/Literatur oder basierend auf Plausibilitätsannahme) → je größer, desto kleineres N benötigt
- **Signifikanzniveau** (i.d.R. alpha = .05) → je kleiner, desto kleineres N benötigt
- **Power** (i.d.R. mindestens 0.8) → je größer, desto kleineres N benötigt

Sampling Plan (Rekrutierungsplanung)

Poweranalyse in R

Poweranalyse für einen unabhängigen t-Test (auch für andere Hypothesentests möglich):

```
> library(pwr)
> pwr.t.test(d = 0.5, sig.level = 0.05, power = 0.80, type = "two.sample", alternative = "greater")
Two-sample t test power calculation

n = 50.1508
d = 0.5
sig.level = 0.05
power = 0.8
alternative = greater

NOTE: n is number in *each* group
```

- Bei a priori Annahme einer Effektstärke von Cohen's d = 0.5 für den Mittelwertsunterschied (mittlerer bis großer Effekt)
- und einem Signifikanzniveau von alpha = .05
- und einer Power von 0.8 (Chance auf positives Ergebnis)

benötigt man N = 102 Personen (51 pro Gruppe), um in einem einseitigen (*alternative = "greater"*) unabhängigen t-Test (*type = "two.sample"*) einen signifikanten Gruppenunterschied nachzuweisen

Sampling Plan (Rekrutierungsplanung)

Sensitivitäts-Poweranalyse in R

Sensitivitäts-Poweranalyse für einen unabhängigen t-Test (auch für andere Hypothesentests möglich):

```
> library(pwr)
> pwr.t.test(d = 0.5, sig.level = 0.05, n = 39, type = "two.sample", alternative = "greater")
Two-sample t test power calculation

n = 39
d = 0.5
sig.level = 0.05
power = 0.7065601
alternative = greater

NOTE: n is number in *each* group
```

- Bei a priori Annahme einer Effektstärke von Cohen's d = 0.5 für den Mittelwertsunterschied (mittlerer Effekt)
- und einem Signifikanzniveau von alpha = .05
- und N = 78 Personen (39 Personen pro Gruppe) (*Hinweis: An der Vorlesung nehmen insgesamt 82 Studierende teil. Wenn die eigene Gruppe = 3-5 Personen nicht an der Studie teilnehmen, aber sonst alle, kommen Sie auf 77-79 Versuchspersonen.*)

erreicht man eine Power von 71% für einen einseitigen, unabhängigen t-Test; d.h. die Wahrscheinlichkeit einen Effekt zu finden, wenn er da ist, beträgt 71%.

Sampling Plan (Rekrutierungsplanung)

Poweranalyse in R: Befehle

- **pwr.t.test()** für Einstichproben-t-test, sowie unabhängige und abhängige Zweistichproben-t-tests mit gleicher Gruppengröße; Angabe von Effektstärke d (s.o.)
- **pwr.anova.test()** für einfaktorielle ANOVA; Angabe von Effektstärke f
 - $f = \sqrt{(\omega^2/1 - \omega^2)}$
 - Angabe von k = Stufen des Faktors
- **pwr.f2.test()** für Regression und mehrfaktorielle ANOVA; Angabe von Effektstärke f^2
 - $f^2 = R^2/(1-R^2)$ bzw. $\omega^2/1 - \omega^2$ für aufgeklärte Varianz des Gesamtmodells
 - $f^2 = (R^2_{AB} - R^2_A)/(1-R^2_{AB})$ bzw. $\omega^2_{partial_B}/1 - \omega^2_{partial_B}$ für aufgeklärte Varianz des Prädiktors B
 - Angabe von u = Anzahl der Prädiktoren/Faktoren ohne Intercept
 - Rückgabewerte v + u + 1 ergibt Stichprobengröße n
- **pwr.r.test()** für Korrelation

Sampling Plan (Rekrutierungsplanung)

Poweranalyse: Woher weiß ich die Effektstärke?

- Recherche in bestehender Literatur! Diese dann herunterkorrigieren (wegen Publication Bias), d.h. erwarteten Effekt für die Poweranalyse kleiner angeben, als das was in der Literatur gefunden wurde.
- Falls nötig, Transformation der Effektstärke aus der Literatur in die Effektstärke, die für die Poweranalyse gebraucht wird (z.B. r in d):
https://www.psychometrica.de/effect_size.html
- Alternativen: Pilotstudien (aber auch nicht zuverlässig) oder praktisch bedeutsame Mindestgrößen
- Bei unklarer zu erwartender Effektstärke: Mehrere Poweranalysen rechnen, visualisieren wie sich die nötige Stichprobenzahl abhängig von der Effektstärke verändert ("Power curves")

14. Transformation of the effect sizes d , r , f , Odds Ratio, η^2 and Common Language Effect Size (CLES) -

Please choose the effect size, you want to transform, in the drop-down menu. Specify the magnitude of the effect size in the text field on the right side of the drop-down menu afterwards. The transformation is done according to Cohen (1988), Rosenthal (1994, S. 239), Borenstein, Hedges, Higgins, and Rothstein (2009; transformation of d in Odds Ratios) and Dunlap (1994; transformation in CLES).

Effect Size	<input type="button" value="d"/> <input type="button" value="0.5"/>
d	0.5
r	0.2425
η^2	0.0588
f	0.25
Odds Ratio	2.4766
Common Language Effect Size CLES	63.82 %
Number Needed to Treat (NNT)	3.6189

Remark: Please consider the additional explanations concerning the transform from d to Number Needed to Treat in the section **BESD and NNT**. When using r as the initial effect size, the calculator draws on the formula specified by Dunlap (1994) for the conversion to CLES: $CLES = \frac{\arcsin(r)}{\pi} + .5$. In all other case d is applied in accordance with McGraw and Wong (1992): $CLES = \frac{d}{\sqrt{2}}$

Sampling Plan (Rekrutierungsplanung)

Poweranalyse: Woher weiß ich die Effektstärke?

- Durchschnittlicher Effekt in der Psychologie: $r \sim .20$ ($d \sim .40$) mit einer Standardabweichung von $r \sim .10\text{-.}15$
- Das ist kleiner als die Einordnung, die Cohen vorgeschlagen hat

Empirical Benchmarks: Psychological Research Literature				
Source	Small	Medium	Large	
Richard et al (2003): characterizations of small, medium and large effects based on 25,000 social psychological studies	.1	.2	.3	
Hemphill (2003): Lower, middle and upper thirds of 380 meta-analyses from psychological assessment and intervention literature	<.2	.2-.3	>.3	
Bosco et al (2015): Lower, middle and upper thirds of 147,328 applied psychology correlations	<.09	.09-.26	>.26	
Gignac & Szodorai (2016): 1st, 2nd, and 3rd quartiles from 87 meta-analyses from individual differences journals	.11	.19	.29	
Paterson et al (2016): 1st, 2nd, and 3rd quartiles of correlations from 258 meta-analyses in organizational behavior	.12	.20	.31	

Cohen's Einordnung:

		small	medium	large
t-test for means	d	.20	.50	.80
t-test for corr	r	.10	.30	.50
F-test for regress	f ²	.02	.15	.35
F-test for anova	f	.10	.25	.40
chi-square	w	.10	.30	.50

Sampling Plan (Rekrutierungsplanung)

Poweranalyse: Woher weiß ich die Effektstärke?

Vergleich mit gut vorstellbaren Effekten, z.B.

- mittlerer Größenunterschied zwischen Männern und Frauen: $d = 1.72$
- mittlerer Placebo-Effekt: $d = 0.24$
- mittlere Veränderung der Lebenszufriedenheit direkt nach Hochzeit: $d = 0.26$
- mittlere Veränderung der Lebenszufriedenheit direkt nach einem Trauerfall: $d = -0.48$
- mittlerer Effekt des Schulbesuchs in der ersten Klasse auf die Mathefähigkeit: $d = 1.1$

Sampling Plan (Rekrutierungsplanung)

Poweranalyse: Wie wähle ich mein Alpha-Niveau?

- Standard in der Psychologie: $\alpha = 0.05$
- Benjamin et al. (2018): Vorschlag, das α -Niveau weiter auf 0.005 zu senken, um Replizierbarkeit der Effekte zu erhöhen
- Lakens et al. (2018): Justify your $\alpha!$
 - Für Ihre Arbeit könnte das bedeuten mal in der Sensitivitätspower-Analyse verschiedene Werte von Alpha auszuprobieren, um zu sehen wie sich die Power dadurch erhöht
 - Meine Empfehlung: Für dieses Ausprobieren α maximal so weit erhöhen wie es nicht den β -Fehler (1-Power) übersteigt
 - Begründung: Im Rahmen dieses Lehrprojekts ist es genauso "schlimm" zu behaupten es gäbe einen Effekt, wenn es keinen gibt, als zu behaupten es gäbe keinen Effekt, wenn es einen gibt.
 - In der Forschung wird der Fehler 1.Art (α -Fehler) zumindest formal konstant niedrig gehalten, während die Höhe des Fehlers 2.Art (β -Fehler) maßgeblich von der Größe der Stichprobe abhängt (welche variabel ist)

Power - Beispiel

Es werden zwei Gruppen mit je 20 Personen als Stichprobe herangezogen ->
Mithilfe von "R" wird also eine Sensitivitäts-Poweranalyse berechnet:

Für die H1 wird ein unabhängiger t-Test berechnet:

```
pwr.t.test (n = 20, d = 0.5, sig.level = 0.05, power = NULL, type = c("two.sample"),  
            alternative = c("greater"))
```

- > da in der Literatur gezeigt wurde, dass sich Frauen deutlich mehr für True Crime Inhalte interessieren, wird von einem mittleren Effekt nach Cohen $d = 0.5$ ausgegangen
- > es ergibt sich eine "power" von 0.4633743. Die Wahrscheinlichkeit, einen Effekt zu finden, wenn er da ist, beträgt also in etwa 46%

Sampling Plan (Rekrutierungsplanung)

Power - Beispiel

Beispiel aus letztem Semester:

Bei der vorliegenden Studie handelt es sich um ein Fixed-N-Design (siehe auch M4). Wir erwarten eine Stichprobengröße von ca. **34 Studienteilnehmer:innen**, die ca. **1:1 auf Experimental- und Kontrollbedingung aufgeteilt** werden. Durch die Randomisierung kann es zu leichten Abweichungen in den Größen der einzelnen Gruppen kommen.

Für unsere Post-Hoc-Sensitivitäts-Power-Analyse legen wir einen **gerichteten unabhängigen t-Test** zugrunde.

Unsere Annahme bezüglich der Effektstärke basiert auf einer Quantifizierung der Effektstärke der von uns geplanten Intervention (MSCI nach Neff, siehe auch M14). Für das Messinstrument SSCS-L wird ein "partial Eta" von 0.428 angegeben (Neff et al., 2021). Da wir die Auswirkungen dieses Effekts der Induktion momentaner Self-Compassion in einem Leistungstest messen und zudem partial Eta tendenziell die Effektgröße überschätzt, reduzieren wir diesen Wert um 50% und nehmen eine **Effektstärke von 0,214** an.

Für einen alpha-Fehler von 0,05 erhalten wir unter den oben beschriebenen Annahmen eine Power von 0,15. Um eine akzeptable Power von 0,8 zu erreichen, müssten wir 542 Teilnehmer:innen rekrutieren. Da dies im Rahmen der Studie nicht möglich ist, entscheiden wir uns bei gegebener Stichprobengröße für einen alpha-Fehler, bei dem alpha- und beta-Fehler gleich wahrscheinlich sind: Für einen **alpha-Fehler von 0,38** erhalten wir eine **Power von 0,62**.

Zum Start der Datenerhebung:

- Alle Datenerhebungen sollten spätestens (!) am 19.06. starten

→ **Damit das klappt, schicken Sie mir Ihren formr run bitte bis Freitag, 19.06. zu.**

Hinweis: Zum Verschicken müssen Sie folgendes in formr auswählen: Run -> Export -> "include survey details" anklicken -> Export. Die exportierte Datei dann an mich schicken.

Literaturempfehlung für die heutige Sitzung

Kapitel 14 und 15 in Döring, N. & Bortz, J. (2016). Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften.
Pearson.