

Diskussionsnotizen OSIP-Meeting 13.11.2019

Anwesende Personen:

OSIP-Mitglieder: Denise Dörfel, Anne Gärtner, Michael Höfler, Veronika Job, Anja Kräplin, Matthias Rudolf, Christoph Scheffel, Martin Schoemann, Ulrike Senftleben, Caroline Surrey
2 Gäste

Protokoll:

Judith Herbers, Denise Dörfel

TOPs:

1. Tagesordnung
2. OpenScienceHack: Umgang mit Extremwerten und Ausreißern
3. Wahl der Open Science Preis Kommission
4. Datenmanagement der Fakultät
5. OSIP Evaluation
6. Vorträge externer Sprecher
7. Planung des nächsten Treffens

1. Tagesordnung

Im heutigen Treffen wird Michael Höfler in einem Vortrag den Umgang mit Ausreißern und Extremwerten beleuchten. Außerdem soll die Kommission für die Vergabe des Open Science Preises bestimmt werden. Die Moderation übernimmt Denise Dörfel.

2. OpenScienceHack: Umgang mit Ausreißern und Extremwerten

Die Folien zum Vortrag von Michael Höfler finden sich unter folgendem Link: https://tu-dresden.de/mn/psychologie/ifap/methpsy/ressourcen/dateien/forschung/openscience/materialien/Hoefler_13112019_-Extremwerte_und_Ausreisser.pdf?lang=de

In Anlehnung an den Vortrag werden die folgenden Aspekte diskutiert:

- Sind in allen Statistiksoftwares robuste Verfahren, wie z.B. die Robuste Regression, implementiert?
 - Nein, leider bietet noch nicht jedes Statistikprogramm direkt die Option einer Robusten Regression, so z.B. nicht SPSS. Möglichkeiten gibt es derzeit aber bereits in Stata, Matlab, Python sowie verschiedenen R-Paketen.
 - Zu beachten ist, dass es aufgrund verschiedener verwendeter Algorithmen zu unterschiedlichen Ergebnissen sowohl zwischen Programmen als auch zwischen den verschiedenen R-Paketen kommen kann.
 - In den neueren Versionen von SPSS sind auch bereits einige robuste Verfahren, z.B. robuste Standardfehler, vorab implementiert. Eine Robuste Regression lässt sich dann (über Umwege) in der SPSS Syntax realisieren. Eine entsprechende Lösungsidee für R und SPSS von Alexander Strobel lautet:

Robuste Regression in R und SPSS

... habe auf Michaels Erklärungen hin, dass robuste Regression die Residuen als Gewichtungsfaktoren nutzt, mal angesehen, ob man das nicht (zumindest als krude Annäherung) mit SPSS umsetzen kann. Zunächst habe ich in R mal ein paar lineare Regressionen bei Daten mit Ausreißern (jetzt mal im allgemeinen Sinne verstanden nicht im Kontrast zu Extremwerten) mit der robusten Regression verglichen. Die lässt sich in R z.B. mit dem Package robustbase und der dort enthaltenen Funktion `lmrob()` umsetzen. Ist ansonsten analog zum `lm()` Syntax für lineare Regression, d.h., statt:

OLS Regression:

```
lm(y ~ x, data)
```

→ Robuste Regression:

```
lmrob(y ~ x, data)
```

Abspeichern der Residuen der `lm()` Prozedur, Bilden des Absolutwerts und Normalisierung am Maximum:

```
res = lm(y ~ x, data)$residuals  
res = 1-(abs(res)/max(abs(res)))
```

gerichtete Regression mit den Residuen als Gewichtungsfaktoren:

```
lm(y ~ x, data, weights = sqrt(1-res))
```

→ fast das identische Resultat wie mit `lmrob()`.

→ Ausreißer somit weniger Einfluss auf die Regressionsgerade (kommt aber mit der obigen Prozedur ggf. auch zu Überkorrektur)

Nutzung dieser Erkenntnis für Vorgehen in SPSS:

→ Regression rechnen und unstandardisierte Residuen abspeichern, ergibt im Data Viewer eine neue Variable (RES_1).

→ Absolutwert bilden

```
(COMPUTE ABSRES_1 = ABS(RES_1). EXECUTE.)
```

→ Über Analysieren>Deskriptive Statistiken Maximum von ABSRES_1 ausgeben lassen (sagen wir, das wäre 15.023439) und kopieren.

→ Normalisieren

```
(COMPUTE NABSRES_1 = 1-(ABSRES_1/15.023439).  
EXECUTE.)
```

→ Wurzel des Resultats bilden

```
(QNABSRES_1 = SQRT(NABSRES_1))
```

(warum auch immer Wurzel bilden, habe gedacht das könnte gut sein, hat in R auch gut funktioniert)

→ Bei der linearen Regression ganz unten QNABSRES_1 ins Feld WLS-Gewichtung schieben.

- Eine in der Praxis oft gesehene Methode ist es, bei Verletzungen der Normalverteilungsannahme die Daten zunächst durch Transformationen annähernd zu normalisieren, sodass im nächsten Schritt weiterhin die Standardverfahren gerechnet werden können. Dies ist z.B. über Box Cox Transformationen möglich. Jedoch führen auch Box Cox Transformationen nur unter gewissen, in der Praxis nicht zwangsläufig erfüllten Voraussetzungen zu normalverteilten Daten, weswegen dieses Vorgehen nur bedingt zu empfehlen ist.

- Sind nicht normalverteilte Daten gegeben und besteht auch keine Annahme über Normalverteilung für die Grundgesamtheit (z.B. bei Reaktionszeitdaten) bietet es sich an, direkt besser geeignete Verteilungen, z.B. Gammaverteilungen, zu nutzen.
- Es ist in einzelnen Fällen bereits vorgekommen, dass die Verwendung robuster Verfahren zu Problemen mit GutachterInnen geführt hat. Eine Erklärung und eine Begründung über die Entscheidung für das robuste Verfahren über das Standardverfahren konnte dem in den meisten Fällen entgegenwirken.

3. Wahl der Open Science Preis Kommission

Für die heutige Sitzung ist die Wahl der Kommission für die Vergabe des Open Science Preises angesetzt. In der letzten Sitzung wurde sich bereits dafür ausgesprochen, die Kommission mit vier Mitgliedern möglichst paritätisch zu besetzen.

Als Kandidaten lassen sich genau vier Mitglieder aufstellen:

- Jens Helmert
- Denise Dörfel
- Martin Schoemann
- Anne Gärtner

Es wird beschlossen, dass Stefan Scherbaum diese Wahlvorschläge per E-Mail mit allen Mitgliedern teilt. Im Falle ausbleibenden Widerspruchs bis zu einem noch festzulegenden Datum gilt die Kommission damit als gewählt.

4. Datenmanagement der Fakultät

Denise Dörfel berichtet, dass unter dem Gesichtspunkt Guter Wissenschaftlicher Praxis (GWP) das Datenmanagement der Fakultät Psychologie überarbeitet werden soll. Eine Kommission bestehend aus den OSIP-Mitgliedern Stefan Scherbaum, Alexander Strobel, Gesine Wieder und Denise Dörfel wird dazu entsprechende Richtlinien entwerfen, um die geforderten GWP Richtlinien einzuhalten. Überarbeitungsbedarf gibt es aktuell beispielsweise bei der Datenhaltung und -sicherung wenn Mitarbeitende den Arbeitsplatz wechseln. Bei der Konzeption kann von bereits implementierten Datenmanagementsystemen anderer Fakultäten und Institutionen, z.B. des Sächsischen Forschungsdatenmanagement oder des ZIH profitiert werden. In diesem Kontext relevant wird auch der Vortrag von Gesine Wieder zum Forschungsdatenmanagementsystem RedCap, der diese Thematik beim nächsten Treffen genauer betrachten wird.

5. OSIP Evaluation

Ulrike Senftleben greift das Thema der OSIP Evaluation aus dem letzten Treffen auf. Sie schlägt vor, gemeinsam mit Judith Herbers ein Konzept für eine zunächst interne OSIP Evaluation zu entwickeln, das dann zu gegebenem Zeitpunkt in einem Treffen vorgestellt werden könnte. Dieser Vorschlag stößt auf breite Zustimmung.

6. Vorträge externer Sprecher

Anja Kräplin merkt an, dass sie aktuell die Vorschläge für externe Sprecher ordnet und bald eine E-Mail an die Mitglieder der Initiative bezüglich der weiteren Koordination schicken wird.

7. Planung des nächsten Treffens

Das nächste OSIP Diskussionstreffen wird am 11.12.2019 wie gewohnt im BZW A307 stattfinden. In diesem wird Gesine Wieder das Forschungsdatenmanagementsystem RedCap vorstellen. Alle Mitglieder sind herzlich eingeladen.