KARDIOTECHNIK 2022; 031(2):000–000
https://doi.org/10.47624/kt.031.TQYX9546

Tutorial

Statistik Teil 2: Datenorganisation mit Tabellenkalkulationsprogrammen

M. Kohl, F. Münch

Die KARDIOTECHNIK wird in der Rubrik Tutorials in Folge relevante Methoden für wissenschaftliche Arbeiten zur klinischen Perfusion vorstellen.

Fazitbox

Pro und Contra Datenorganisation mit Tabellenkalkulationsprogrammen:

Pro
- Tabellenkalkulationsprogramme eignen sich in vielen Fällen sehr gut zur Eingabe, Organisation und Speicherung von Daten.
- Werden die Daten adäquat organisiert, erleichtert und beschleunigt dies eine fehlerfreie statistische Analyse.

Contra
- Tabellenkalkulationsprogramme stoßen bei komplexen oder sehr umfangreichen Daten („big data“) an ihre Grenzen.
- Datenredundanz ist üblicherweise nicht vermeidbar.

Der Ausgangspunkt für eine effiziente und fehlerfreie statistische Analyse ist eine adäquate Erhebung, Organisation und Speicherung der Daten. Dabei sollte das Ziel eines guten Datenmanagements immer sein, die FAIR-Prinzipien („findable, accessible, interoperable, reusable“) einzuhalten [1]. Darüber hinaus sollten personenbezogene Daten immer in pseudonymisierter Form abgespeichert werden; d. h., es muss sichergestellt sein, dass die Daten nicht mehr einer konkreten Person zugeordnet werden können (vgl. § 46 Abs. 5 des Bundesdatenschutzgesetzes, https://www.buzer.de/46_BDSG_Bundesdatenschutzgesetz.htm).

Die folgenden Empfehlungen zur Organisation von Daten mittels Tabellenkalkulationsprogrammen (TKPs) basieren zum größten Teil auf den Arbeiten von Broman und Woo [2], Wickham [3] und White et al. [4]. Diese Referenzen enthalten außerdem viele hilfreiche Beispiele.

Zunächst einmal sollte man beachten, dass die Datenorganisation bereits vor der Datenerhebung beginnt, da das Umorganisieren von bereits erhobenen Daten eine bekannte Fehlerquelle ist. Sollte dies wirklich nötig werden, so empfiehlt es sich, die Umorganisation mit Hilfe von einem eigens dafür geschriebenen, gut dokumentierten Code durchzuführen. So kann man die Änderungen jederzeit nachvollziehen, prüfen und gegebenenfalls anpassen [2]. Sehr gut eignen sich dafür etwa die Statistiksoftware R [5] (z. B. R Skript oder R Markdown Dokument [6]) oder die Programmiersprache Python [7] (z. B. Python Skript oder Jupyter Notebook [8]). Weiterhin ist es von enormer Wichtigkeit, dass man bei der Datenorganisation bezüglich aller Aspekte (Ordnerstruktur, Ordnernamen, Dateinamen, Tabellenaufbau, Variablennamen, Variablenwerte, etc.) möglichst konsistent vorgeht [2].

Werden die Daten mit Hilfe eines TKPs erhoben und gespeichert, sollten für jeden Datensatz zwei Tabellen angelegt werden. Die erste Tabelle dient hierbei zur Aufnahme der erhobenen Daten, die zweite Tabelle enthält das sogenannte Datenwörterbuch („data dictionary“). Die Datentabelle sollte dabei so organisiert werden, dass sie die erhobenen Variablen als Spalten enthält, wobei die erste Zeile die Spaltennamen enthalten sollte (Tab. 1–3). Die Zeilen hingegen entsprechen den Beobachtungen (z. B. Patienten), wobei es auch möglich ist, in der ersten Spalte Zeilennamen zu vergeben, falls dies sinnvoll erscheint [2,3,4].

Wickham bezeichnet einen derartigen Aufbau als „tidy data“ (aufgeräumte Daten) [3]. Bei der Vergabe von Zeilen- und Spaltennamen ist insbesondere darauf zu achten, dass diese für sich genommen eindeutig sind, d. h. weder innerhalb der Zeilen- noch innerhalb der Spaltennamen sollte ein Name wiederholt auftauchen. Die Variablennamen (= Spaltennamen) sollten kurz, prägnant und aussagekräftig sein, wobei man auf die Verwendung von Leer- und Sonderzeichen (z. B. #, @, €, ° etc.) sowie Umlaute und ß verzichten sollte [2] (Tab.1–3).

Die Zellen der Datentabelle enthalten die erhobenen Werte, wobei im Fall von kategorialen (nominal- oder ordinalskalierten) Variablen die Namen der Kategorien ebenfalls nach obigen Regeln gewählt werden sollten. Generell empfiehlt es sich, Werte unter Einhaltung internationaler Standards zu wählen; so sollte z. B. für Datumswerte das Format „YYYY-MM-DD“ nach ISO 8601 verwendet werden (z. B. 19. Februar 2022 = 2022-02-19) [2,4]. Ein weiterer wichtiger Aspekt beim Erheben von Daten ist die Kodierung von fehlenden Werten. Generell sollte eine Datentabelle nach der Datenerhebung keine leeren Zellen geschweige denn leere Zeilen oder Spalten enthalten. Im Fall eines fehlenden Wertes sollte daher die entsprechende Zelle mit einem speziellen Wert gefüllt werden [2]. Die Abkürzung NA („not available“) ist in den allermeisten Fällen eine gute Option [2,4].

Im Folgenden finden sich in kurzer Form weitere wichtige Empfehlungen zur Datenorganisation [2,3,4]:
- *Jede Zelle sollte nur eine einzelne Information enthalten.* Wird etwa nur unter bestimmten Bedingungen eine Messung durchgeführt, so sollte man eine Spalte anlegen, in der verzeichnet ist, ob gemessen wurde und eine zweite Spalte, welche die Messwerte enthält.
- *Daten und darauf basierende Berechnungen sollten strikt getrennt werden,* d. h. in der Datentabelle sollten keine, auch nicht einfache, Berechnungen durchgeführt werden.
- *Auf eine Formatierung von Zellen* (Rahmen, Farben, Schattierung, etc.) *oder der Schriftart* (farbig, fett, kursiv, etc.) *sollte verzichtet werden.* Dienen Formatierungen speziell dazu, zusätzliche Information darzustellen, z. B. ob es sich um einen validierten Wert handelt oder nicht, so sollte diese Information in einer zusätzlichen Spalte abgespeichert werden.

Prof. Dr. Matthias Kohl
Department of Medical and Life Sciences
Institute of Precision Medicine
Hochschule Furtwangen
Jakob-Kienzle-Str. 17,
78054 Villingen-Schwenningen (Germany)
E-Mail: kohl@hs-furtwangen.de
www.hs-furtwangen.de
www.life-data-science.org

ID	Datum	Gruppe	Zeitpunkt	Kalium	Laktat
1	2011-04-04	Kontrolle	T0	4,10	1,60
6	2011-04-15	Intervention	T0	4,00	1,80
10	2011-05-05	Kontrolle	T0	3,90	2,60
19	2011-05-27	Intervention	T0	3,80	1,90
27	2011-08-26	Intervention	T0	3,20	2,40
34	2011-09-05	Kontrolle	T0	6,10	2,60
38	2011-10-11	Intervention	T0	3,70	2,10
42	2011-11-10	Intervention	T0	3,10	1,20
46	2011-11-29	Kontrolle	T0	3,50	2,50
47	2011-11-30	Kontrolle	T0	3,50	1,40
1	2011-04-04	Kontrolle	T1	7,50	9,50
6	2011-04-15	Intervention	T1	4,90	10,40
10	2011-05-05	Kontrolle	T1	7,10	8,60
19	2011-05-27	Intervention	T1	5,60	11,60
27	2011-08-26	Intervention	T1	6,50	7,50
34	2011-09-05	Kontrolle	T1	7,90	8,00
38	2011-10-11	Intervention	T1	4,90	8,40
42	2011-11-10	Intervention	T1	NA	6,80
46	2011-11-29	Kontrolle	T1	5,60	8,70
47	2011-11-30	Kontrolle	T1	NA	NA

Tab. 1a: Einfache Datentabelle in langem Format.

Variable	Beschreibung	Werte
ID	Tier-ID	Fortlaufend beginnend mit 1
Datum	Datum	Datum in Form YYYY-MM-DD
Gruppe	Studiengruppe	Kontrolle = Kontrollgruppe; Intervention = Interventionsgruppe
Zeitpunkt	Zeitpunkt der Messung	T0, T1
Kalium	Kaliumwerte	Kalium in mmol/l
Laktat	Laktatwerte	Laktat in mmol/l

Tab. 1b: Datenwörterbuch in langem Format.

– *Es sollten regelmäßig Backups von den Datendateien durchgeführt werden.* Eine alternative Möglichkeit zur Sicherung der Daten stellen auch sogenannte öffentliche Datenspeicher („data repositories“) dar.
– *Die Daten sollten regelmäßigen Qualitätskontrollen unterzogen werden.* Einfache Beispiele sind etwa: Haben alle Spalten den richtigen Datentyp; liegen alle Werte innerhalb plausibler Wertebereiche; finden sich nicht-numerische Werte in einer Spalte, die nur numerische Werte enthalten sollte?
– *Die Daten verschiedener Beobachtungseinheiten sollten nicht vermischt werden.* Werden z. B. in einer multizentrischen Studie zum einen Daten über die beteiligten Kliniken und zum anderen über die einzelnen Patienten gesammelt, so sollten die Daten zu den Kliniken getrennt von den Daten der Patienten abgespeichert werden.
– *Es ist besser, für jede Tabelle eine eigene Datei zu erstellen, anstelle nur eine Datei mit mehreren Tabellenblättern zu verwenden.*
– *Es ist am besten, die Tabellen in einfachem Textformat zu speichern* (Spalten z. B. durch Strichpunkt oder Tabulator getrennt). Dies stellt zum einen sicher, dass für die Dateien nie eine spezielle Software benötigt wird, und zum anderen kann jedes TKP diese problemlos öffnen und in der üblichen Form anzeigen.

Die Variablennamen, die möglichen Werte oder Wertebereiche der Variablen sowie eventuelle weitere Erklärungen werden als Zeilen in das Datenwörterbuch eingetragen.

In Tabelle 1 findet sich eine einfache Datentabelle mit dem zugehörigen Datenwörterbuch. In dieser sogenannten langen Form lassen sich viele statistische Analysen oder graphische Darstellungen direkt, ohne zusätzliche Transformationen, durchführen.

Die Daten können aber auf Basis der obigen Empfehlungen auch auf andere Weise organisiert werden. In Tab. 2 findet sich eine Alternative, die auch als breite Form bezeichnet wird. In diesem Fall werden die Messwerte der verschiedenen Zeitpunkte in separaten Spalten aufgeführt. Diese Form der Datenorganisation ist zum Beispiel vorteilhaft, wenn in der statistischen Analyse die Werte eines Zeitpunktes auf die Werte eines anderen Zeitpunktes adjustiert werden sollen (z. B. im Fall der ANCOVA).

Die dritte Tabelle schließlich zeigt eine weitere Möglichkeit, diesen Datensatz zur organisieren, sozusagen eine „extra lange“ Form (Tab. 3). In diesem Fall werden alle Messwerte in einer einzigen Spalte erfasst, was eine weitere Spalte nötig macht, in der festgehalten ist, um welchen Analyten es sich handelt. Auch diese Form eignet sich gut für nachfolgende statistische Analysen oder graphische Darstellungen.

Außerdem kann mit geeigneter Software wie z. B. R [5], Python [7] oder auch IBM SPSS Statistics recht einfach zwischen langen und breiten Formaten hin und her transformiert werden.

Organisiert man die Daten wie oben beschrieben, eignen sich TKPs in vielen Fällen sehr gut für die Eingabe, Organisation und Speicherung von Daten und ermöglichen eine effiziente und fehlerfreie statistische Analyse. Jedoch sollte man sich der Schwächen bewusst sein. Die von der European Spreadsheet Risks Interest Group (EuSpRiG) gesammelten Fälle zeigen Fehler auf und wie diese hätten vermieden werden können [9]. So hat etwa die Verwendung des xls-Formats mit seiner Beschränkung auf ca. 65.000 Zeilen im September 2020 dazu geführt, dass nahezu 16.000 COVID-19 Fälle in Großbritannien undokumentiert blieben [10]. Weiterhin ist seit mehreren Jahren bekannt, dass die Autokorrektur von Microsoft Excel regelmä-

ID	Datum	Gruppe	Kalium_T0	Kalium_T1	Laktat_T0	Laktat_T1
1	2011-04-04	Kontrolle	4,10	7,50	1,60	9,50
6	2011-04-15	Intervention	4,00	4,90	1,80	10,40
10	2011-05-05	Kontrolle	3,90	7,10	2,60	8,60
19	2011-05-27	Intervention	3,80	5,60	1,90	11,60
27	2011-08-26	Intervention	3,20	6,50	2,40	7,50
34	2011-09-05	Kontrolle	6,10	7,90	2,60	8,00
38	2011-10-11	Intervention	3,70	4,90	2,10	8,40
42	2011-11-10	Intervention	3,10	NA	1,20	6,80
46	2011-11-29	Kontrolle	3,50	5,60	2,50	8,70
47	2011-11-30	Kontrolle	3,50	NA	1,40	NA

Tab. 2a: Einfache Datentabelle in breitem Format.

Variable	Beschreibung	Werte
ID	Tier-ID	Fortlaufend beginnend mit 1
Datum	Datum	Datum in Form YYYY-MM-DD
Gruppe	Studiengruppe	Kontrolle = Kontrollgruppe; Intervention = Interventionsgruppe
Kalium_T0	Kaliumwerte zum Zeitpunkt T0	Kalium in mmol/l
Kalium_T1	Kaliumwerte zum Zeitpunkt T1	Kalium in mmol/l
Laktat_T0	Laktatwerte zum Zeitpunkt T0	Laktat in mmol/l
Laktat_T1	Laktatwerte zum Zeitpunkt T1	Laktat in mmol/l

Tab. 2b: Datenwörterbuch in breitem Format.

ßig Gennamen verfälscht und ca. 30 % der publizierten Artikel fehlerhafte Gennamen in den ergänzenden Daten enthalten, weshalb mittlerweile sogar einige Gensymbole umbenannt wurden [11]. Auch ist bei der Datenorganisation mit einem TKP, wie im obigen Beispiel zu sehen ist, das Prinzip der Vermeidung von Datenredundanz üblicherweise verletzt. Bei komplexen oder sehr umfangreichen Daten („big data“) empfiehlt sich die Entwicklung einer entsprechend angepassten und optimierten Datenbank.

LITERATUR

[1] Wilkinson M, Dumontier M, Aalbersberg I et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3. 2016; 160018. doi: 10.1038/sdata.2016.18/.

[2] Broman KW, Woo KH. Data Organization in Spreadsheets. The American Statistician. 2018; 72:1,2-10. doi: 10.1080/00031305.2017.1375989/.

[3] Wickham H. Tidy Data. Journal of Statistical Software. 2014; 59(10), 1-23. doi: 10.18637/jss.v059.i10/.

[4] White EP, Baldridge E, Brym ZT, Locey KJ, McGlinn DJ, Supp, SR. Nine simple ways to make it easier to (re) use your data. Ideas in Ecology and Evolution. 2013; 6(2).

[5] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2022; Vienna, Austria. URL https://www.R-project.org/.

[6] Xie Y, Dervieux C, Riederer E. R Markdown Cookbook. Chapman and Hall/CRC. 2020; ISBN 9780367563837. URL https://bookdown.org/yihui/rmarkdown-cookbook.

[7] Python Software Foundation. Python. 2022; URL https://www.python.org/, letzter Zugriff 19.02.2022.

[8] Beg M et al. *(2021). Using Jupyter for Reproducible Scientific Workflows.* Computing in Science & Engineering. *2021; 23(2), 36-46. doi: 10.1109/MCSE.2021.3052101/.*

[9] O'Beirne P. European Spreadsheet Risk Interest Group. EuSpRiG Horror Stories: Spreadsheet-mistakes – news stories. http://www.eusprig.org/horror-stories.htm, letzter Zugriff 19.02.2022.

[10] Kelion L. Excel: Why using Microsoft's tool caused Covid-19 results to be lost. 2020. https://www.bbc.com/news/technology-54423988, letzter Zugriff 19.02.2022.

[11] Lewis D. Autocorrect errors in Excel still creating genomics headache. Nature NEWS. 2021. doi: 10.1038/d41586-021-02211-4/.