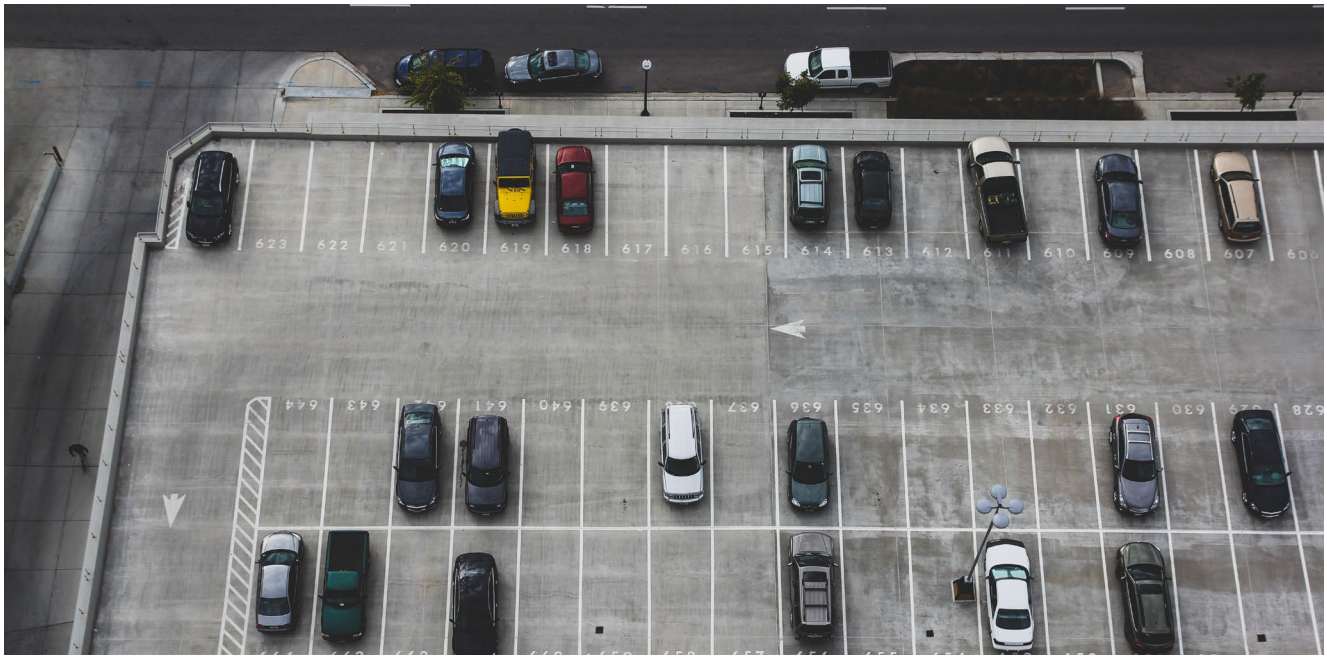


© John Matychuk on Unsplash

Was tun, wenn ich keine Normalverteilung habe?

von Daniela Keller

In meinem letzten Exposé-Beitrag[1] habe ich Ihnen von der Normalverteilung erzählt: was die Normalverteilung eigentlich ist, warum sie in den statistischen Analysen so wichtig ist und wie Sie sie prüfen können.

Nun stelle ich mir vor, Sie haben die Normalverteilung für Ihre Daten untersucht, weil Sie eine statistische Methode rechnen möchten, die als Voraussetzung Normalverteilung fordert, beispielsweise einen t-Test, eine mehrfaktorielle ANOVA oder eine lineare Regression. Leider sind Ihre Daten aber nicht normalverteilt. Was nun?

In diesem Artikel stelle ich Ihnen fünf Wege vor, die Sie in dieser Situation einschlagen können.

1. Ausreißer bereinigen

Wenn das eigentliche Problem gar nicht die fehlende Normalverteilung ist, sondern Ihre Daten bis auf einzelne Ausreißer annähernd normalverteilt sind, dann dürfen Sie diese Ausreißer bereinigen und im Anschluss mit den Daten ohne diese extremen Messwerte weiterarbeiten. Sie sollten dieses Vorgehen sehr genau dokumentieren, damit Ihre Schritte für die Leser*innen Ihrer Arbeit am Ende nachvollziehbar sind. Und Achtung: Dieser Weg führt nur dann zum Ziel, wenn es sich um *einzelne* extreme Werte handelt und durch Löschen dieser Werte nicht neue Ausreißer entstehen!

2. Daten transformieren

Liegt der Mangel an Normalverteilung daran, dass die Daten eine deutliche Schiefe aufweisen, dann kann es hilfreich sein, sich auf die Suche nach einer geeigneten Transformation zu machen. Eine Datentransformation ist die Anwendung einer Formel (z. B. des Logarithmus) auf Ihre Daten. Durch diese Formel wird die Verteilung der Daten geändert. Im Idealfall sind die transformierten Daten dann (besser) normalverteilt. So eine Transformation ist keine Datenmanipulation und ist erlaubt, wenn

1 https://www.budrich-journals.de/index.php/expose/article/view/40914

Sie sich an die Regel halten, keine Formel zu verwenden, die die Reihenfolge der Daten ändert. Es dürfen nur die Abstände zwischen den Messwerten geändert werden, aber nicht deren Reihenfolge. Das ist für alle gängigen Transformationen wie z. B. die Wurzelfunktion, den Logarithmus, den Kehrwert oder auch die Tukey- oder Box-Cox-Transformationen der Fall.

Wenn Sie eine geeignete Transformation finden, rechnen Sie die schließenden Statistiken mit den transformierten Daten weiter. Auch hier müssen Sie wieder gut dokumentieren, was Sie tun, damit Ihre Analyse transparent bleibt.

3. Über Robustheit argumentieren

Bei einigen statistischen Methoden und wenn Ihr Datensatz nicht zu klein ist, können Sie trotz verletzter Normalverteilungsannahme die parametrische Methode (= Methode, die Normalverteilung braucht) rechnen, da Sie wissen, dass sie robust auf Abweichungen von der Normalverteilung reagiert. Das ist beispielsweise für die Varianzanalyse der Fall, wenn in jeder Gruppe mindestens zehn Beobachtungen liegen (Bortz, 2005).

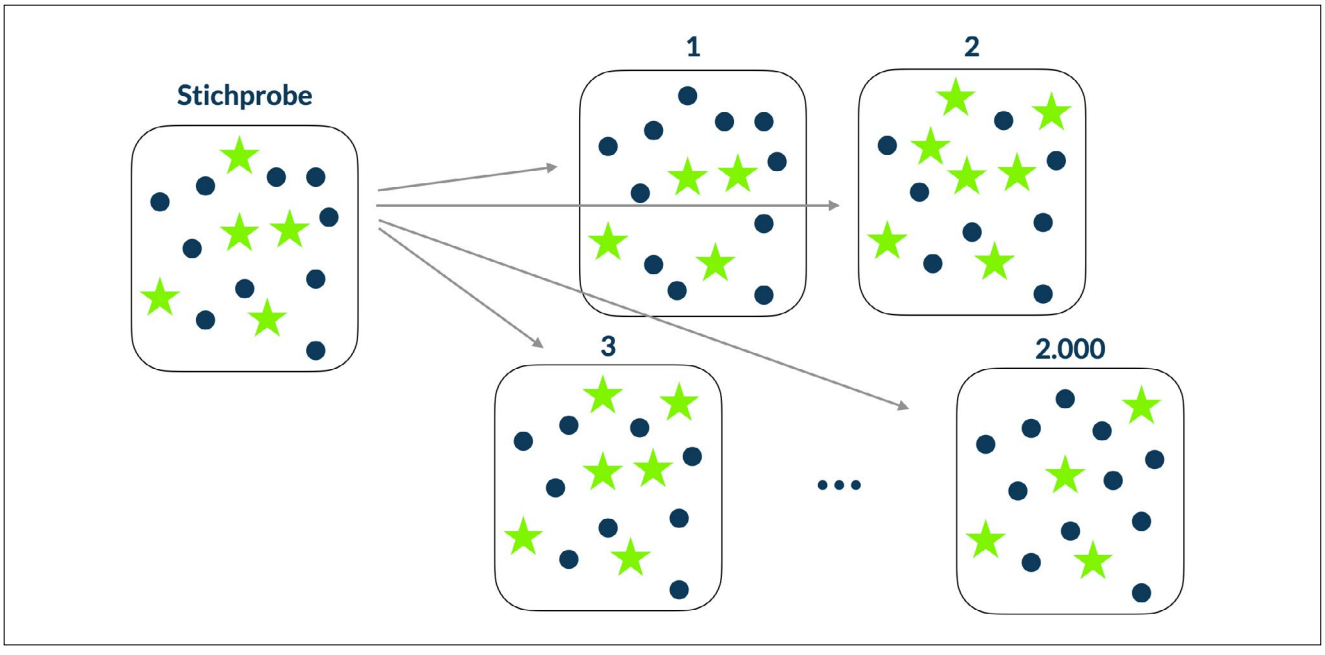
Wenn Ihr Datensatz nicht sehr klein ist, dann schauen Sie also, ob Sie über die Robustheit der von Ihnen geplanten Methode Informationen finden. Diese nutzen Sie dann, um trotz nicht erfüllter Normalverteilungsannahme die Methode einsetzen zu können.

4. Nicht-parametrische Methoden verwenden

Für viele parametrische Methoden stehen nicht-parametrische Alternativen zur Verfügung. Das gilt vor allem für die einfacheren Methoden, bei denen nur mit einer oder zwei Variablen gerechnet wird, beispielsweise für die bivariaten Korrelationen oder die Tests auf Lageunterschiede. Hier gibt es immer auch einen Signifikanztest, der keine Normalverteilung benötigt bzw. für den es egal ist, ob die Daten normalverteilt sind oder nicht.

Beispielsweise ist der Mann-Whitney-U-Test die nicht-parametrische Alternative des t-Tests für unabhängige Stichproben, für den t-Test für gepaarte Stichproben wird der Wilcoxon-Test eingesetzt und statt der einfaktoriellen ANOVA der Kruskal-Wallis-Test. Für die Pearson-Korrelation können Sie im nicht-parametrischen Fall die Spearman- oder Kendall-Korrelation verwenden.

Diese nicht-parametrischen Methoden arbeiten statt mit den Messwerten mit den Rängen der Daten und sind dadurch zusätzlich robust gegenüber Ausreißern. Die nicht-parametrischen Methoden haben keine Nachteile gegenüber den parametrischen. Allein, falls doch Normalverteilung vorläge, hätten Sie eine etwas höhere



Teststärke und könnten in diesem Fall einen Effekt leichter als signifikant nachweisen. Wenn aber eben keine Normalverteilung besteht, sind die nicht-parametrischen Methoden das Mittel der Wahl.

5. Bootstrapping nutzen

Wenn Sie mit einem größeren Datensatz arbeiten (mindestens $N = 50$) und Ihre Statistiksoftware Bootstrapping im Angebot hat, dann können Sie das nutzen, um sich gegenüber einer verletzten Normalverteilungsannahme abzusichern. In diesem Fall rechnen Sie die Analyse wie geplant (obwohl eben die Normalverteilungsvoraussetzung nicht erfüllt ist) und rechnen zusätzlich ein Bootstrapping. Im Bootstrapping wird aus Ihrem Datensatz sehr, sehr oft (z. B. 2.000 Mal) ein neuer Datensatz der gleichen Stichprobengröße mit Zurücklegen gezogen. Dann wird auf diesen sehr, sehr vielen neuen Datensätzen jeweils die gewünschte Analyse gerechnet und diese vielen Ergebnisse dann zu einem Gesamtergebnis zusammengefasst. Das ist das Bootstrapping-Ergebnis. Dieses Bootstrapping-Ergebnis ist verlässlich, selbst wenn die Voraussetzung der Normalverteilung der Daten nicht erfüllt war.

Nun haben Sie also verschiedene Möglichkeiten an der Hand, um das Problem mit der fehlenden Normalverteilung zu umgehen. Sie sehen, dass es sehr unterschiedliche Möglichkeiten sind, und oft passen in einer speziellen Situation auch nur eine oder wenige davon. Schauen Sie also, was genau bei Ihnen Sache ist:

- Was führt zur Nicht-Normalverteilung? Ausreißer, Schiefe…
- Gibt es nicht-parametrische Alternativen für Ihre Analysemethode?
- Wie groß ist Ihr Datensatz?
- Was kann Ihre Software?

Wenn Sie diese Fragen geklärt haben, finden Sie leicht den in Ihrer Situation passenden Weg und können Ihre statistische Auswertung verlässlich und korrekt fortsetzen.

Referenz

Bortz, Jürgen (2005): Statistik für Human- und Sozialwissenschaftler. 6. Auflage. Heidelberg: Springer, S. 287.

© privat

Die Autorin

Daniela Keller ist leidenschaftliche Statistik-Expertin und berät Studierende und Wissenschaftler*innen zu allen Themen der statistischen Datenanalyse. Während ihres Studiums der Diplom-Mathematik gründete sie mit Kommilitonen eine studentische statistische Beratung und arbeitete anschließend selbständig in diesem Feld. Neben Einzelberatungen und Workshops unterstützt sie Ihre Kund*innen seit 2019 mit der Statistik-Akademie, ihrem Online-Mitgliederbereich für alle, die Statistik verstehen und selbständig anwenden wollen. Ihr Blog (www.statistik-und-beratung.de/blog) und ihr YouTube-Kanal sind Fundgruben für leicht verständlich aufbereitetes Statistikwissen für die Praxis.